Conteo de Bacterias

Gómez Jiménez Aaron Mauricio - Francisca Fernanada Vilca Sánchez

2023-05-09

Un investigador está interesado en determinar el log_{10} de conteos de microbios obtenidos de un cupó contaminado de 2.3 cm^2 a diferentes temperaturas y en diferentes medios. La hipótesis sostiene que las variaciones de temperatura de 20.8 žC a 40.8 žC y la concentración del medio afectarían los conteos. ¿Puede apoyar la hipótesis con un modelo lineal?

Lo primero hacer es cargar los datos, renombrar las variables y se realiza un análisis exploratorio con las estadísticas más comunes.

```
summary(datos)
```

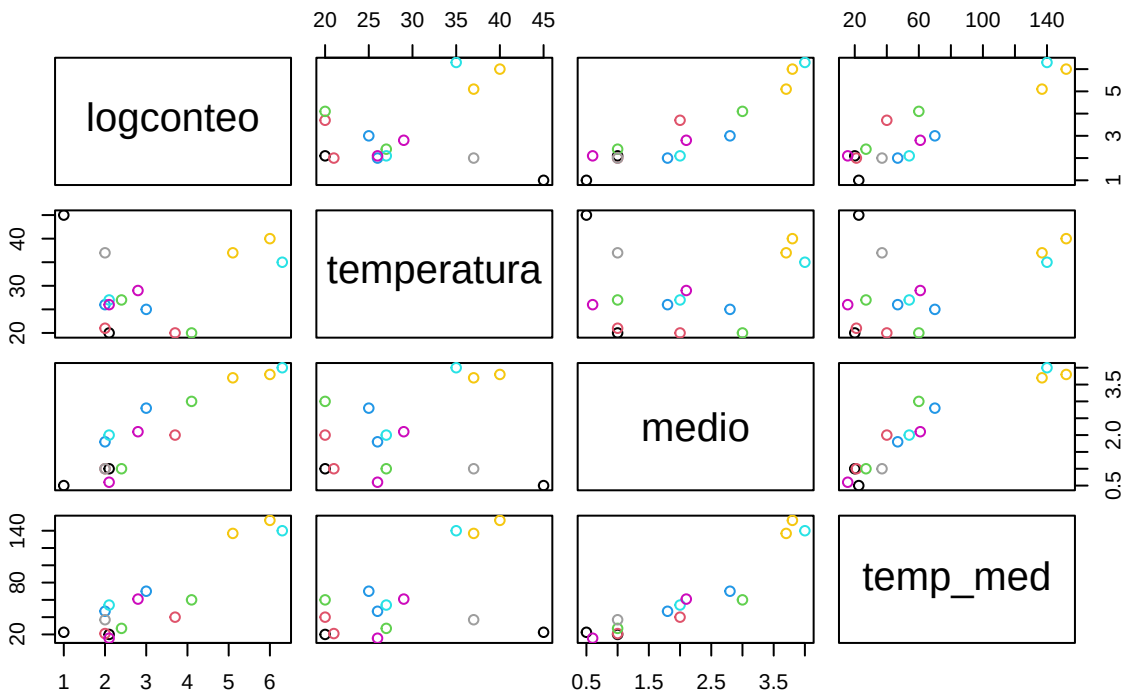
```
##   observación     logconteo      temperatura     medio         temp_med
##  Min.   : 1.0   Min.   :1.000   Min.   :20   Min.   :0.50   Min.   : 15.60
##  1st Qu.: 4.5   1st Qu.:2.050   1st Qu.:23   1st Qu.:1.00   1st Qu.: 24.75
##  Median : 8.0   Median :2.400   Median :27   Median :2.00   Median : 46.80
##  Mean   : 8.0   Mean   :3.113   Mean   :29   Mean   :2.02   Mean   : 60.25
##  3rd Qu.:11.5   3rd Qu.:3.900   3rd Qu.:36   3rd Qu.:2.90   3rd Qu.: 65.45
##  Max.   :15.0   Max.   :6.300   Max.   :45   Max.   :4.00   Max.   :152.00
```

Visualizando los datos y calculando la correlación entre las variables, se puede ver si tiene sentido hacer la regresión

```
##             logconteo temperatura     medio  temp_med
## logconteo   1.0000000   0.2144943 0.9199902 0.9134577
## temperatura 0.2144943   1.0000000 0.1872706 0.4648337
## medio       0.9199902   0.1872706 1.0000000 0.9356600
## temp_med    0.9134577   0.4648337 0.9356600 1.0000000
```

Notamos que existe una correlación alta entre el `conteo de bacterias` y el `medio`, también entre la `temperatura*medio` y el `conteo de bacterias`.

```
plot(datos[,2:5], col=palette("ggplot2"), main="Pairs de los Datos")
```



Al graficar los datos por pares, notamos que podría existir una relación lineal entre el `conteo` y el `medio`, puesto que, en su gráfica podemos ver una tendencia positia y lineal, lo mismo podemos concluir de el `conteo` y de la `temperatura*medio`, más adelante se analizará más a fondo este caso.

Haciendo una regresión para explicar si el número de bacterias depende de la temperatura

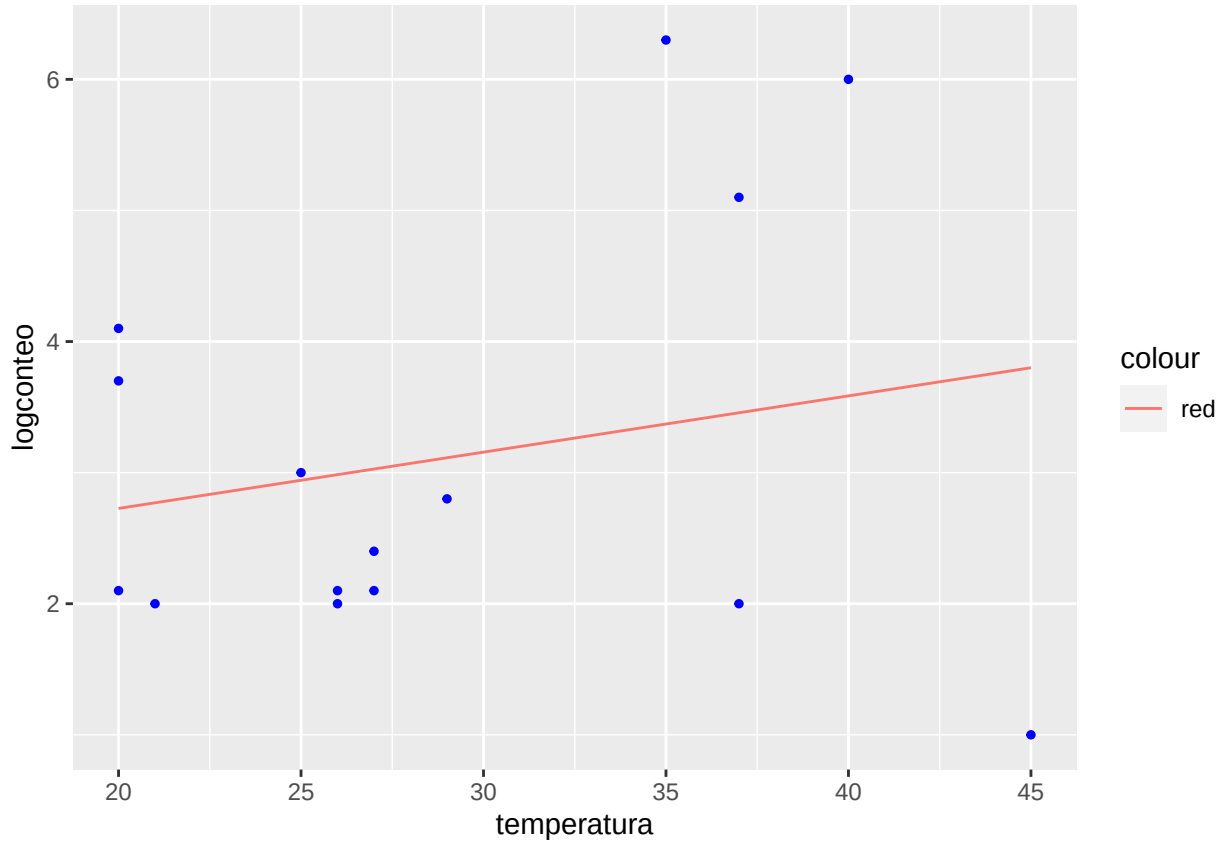
```
reg1=lm(logconteo ~ temperatura, data=datos)
summary(reg1)
```

```
##
## Call:
## lm(formula = logconteo ~ temperatura, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.800 -0.906 -0.627  1.173  2.929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.86861    1.62652   1.149    0.271
## temperatura  0.04292    0.05421   0.792    0.443
##
## Residual standard error: 1.617 on 13 degrees of freedom
## Multiple R-squared:  0.04601,    Adjusted R-squared:  -0.02738
## F-statistic: 0.6269 on 1 and 13 DF,  p-value: 0.4427
```

Al analizar la variable explicativa queremos saber si es significativa, como obtenemos un p-value 0.443 usando

una significancía de 0.05, concluimos que no se rechaza la hipotesis nula, es decir podría ser $B_0 = 0$, ademas de una R cuadrada muy baja de 0.04601, es decir con nuestra regresión se explican solo el 4% de los datos.

```
scatter_line = ggplot(datos, aes(x = temperatura, y = logconteo)) + geom_point(color = "blue", size = 1
scatter_line + geom_line(aes(x = temperatura, y = reg1$fitted.values, color = "red"))
```



Ahora haciendo la regresión para el conteo y el medio

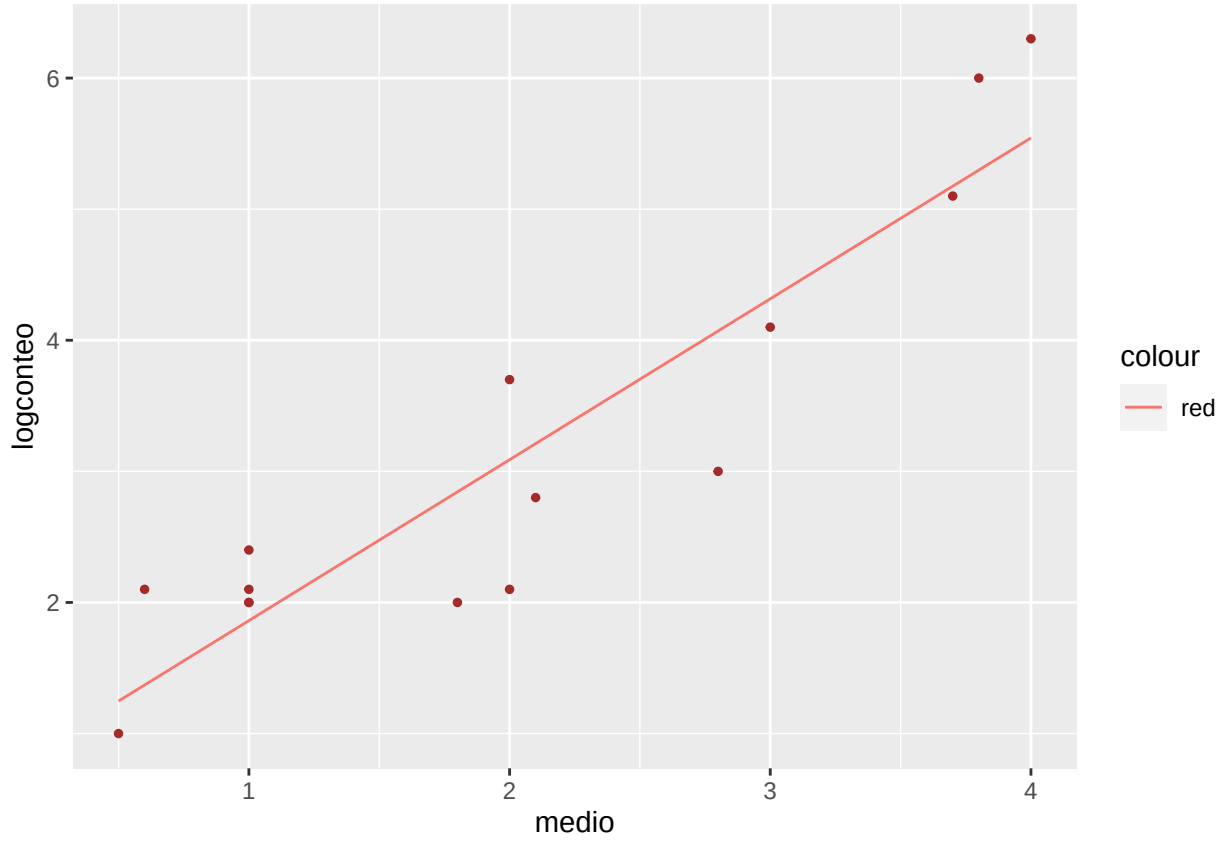
```
reg2=lm(logconteo~medio, data=datos)
summary(reg2)
```

```
##
## Call:
## lm(formula = logconteo ~ medio, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0707 -0.3297  0.1385  0.5749  0.7566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6341     0.3375   1.879   0.0828 .
## medio         1.2273     0.1450   8.463  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.6489 on 13 degrees of freedom
## Multiple R-squared:  0.8464, Adjusted R-squared:  0.8346
## F-statistic: 71.63 on 1 and 13 DF,  p-value: 1.201e-06
```

En esta regresión podemos observar que la variable explicativa medio es significativa con el alfa usual, y ademas obetenemos una R cuadrada muy buena de 0.8464, por lo cual podemos asegurar que *existe una relación lineal positiva.*

```
scatter_line = ggplot(datos, aes(x = medio, y = logconteo)) + geom_point(color = "brown", size = 1)
scatter_line + geom_line(aes(x = medio, y = reg2$fitted.values, color = "red"))
```



Ahora para explicar el conteo con la multiplicación de la temperatura y el medio

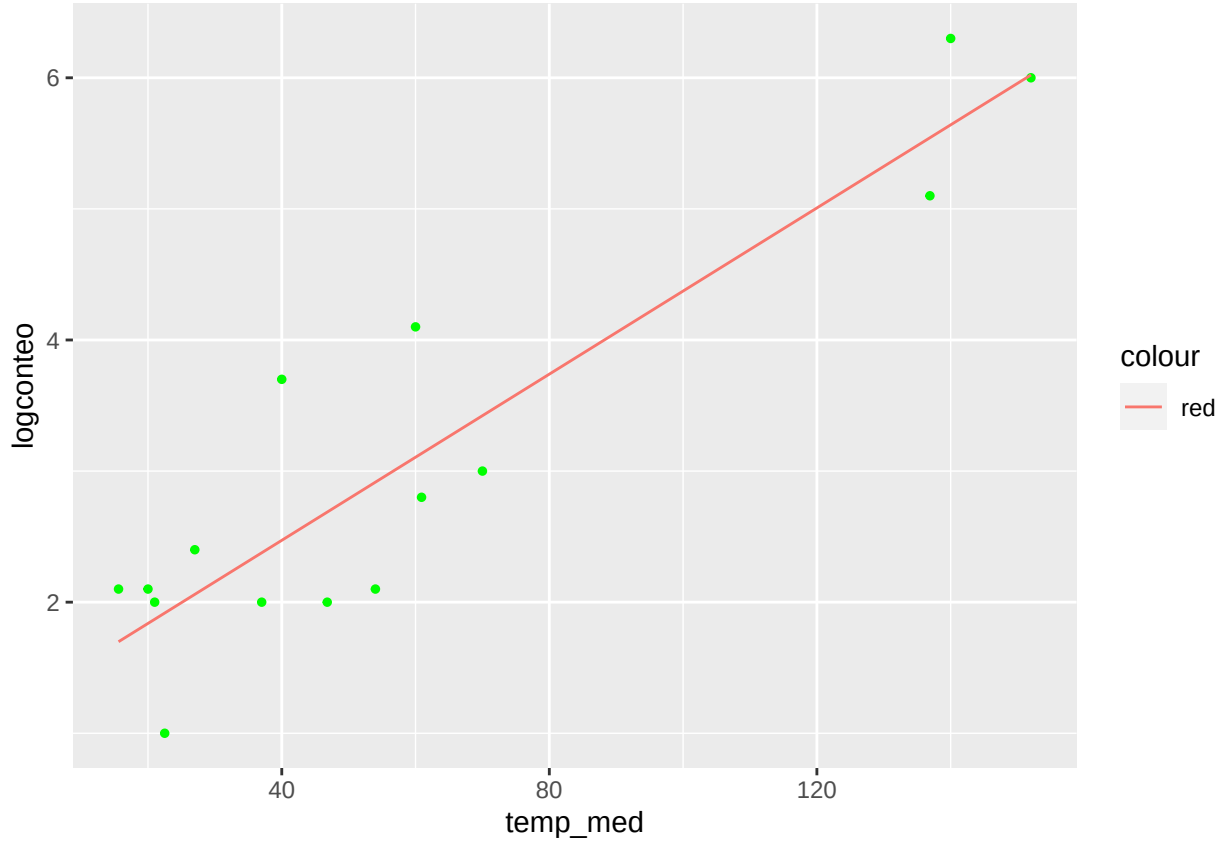
```
reg3=lm(logconteo~ temp_med, data=datos)
summary(reg3)
```

```
##
## Call:
## lm(formula = logconteo ~ temp_med, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91703 -0.43258 -0.02127  0.37100  1.22834
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.203936   0.293122   4.107  0.00124 **
## temp_med    0.031693   0.003916   8.094 1.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6738 on 13 degrees of freedom
## Multiple R-squared:  0.8344, Adjusted R-squared:  0.8217
## F-statistic:  65.5 on 1 and 13 DF,  p-value: 1.969e-06
```

Observamos que la combinación lineal de las variables explicativas es significante, ademas que nos da una R cuadrada de .83 es decir se explican el 83% de los datos. Por lo tanto para nuestro siguiente modelo incluiremos el medio y la temp*medio.

```
scatter_line = ggplot(datos, aes(x = temp_med, y = logconteo)) + geom_point(color = "green", size = 1)
scatter_line + geom_line(aes(x = temp_med, y = reg3$fitted.values, color = "red"))
```




```
reg4=lm(logconteo~ medio+ temp_med, data=datos)
summary(reg4)
```

```
##
## Call:
## lm(formula = logconteo ~ medio + temp_med, data = datos)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9077 -0.4376  0.1760  0.3963  0.8977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.81643    0.34904   2.339   0.0375 *
## medio        0.69953    0.39550   1.769   0.1023
## temp_med     0.01467    0.01029   1.426   0.1793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6246 on 12 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.8468
## F-statistic: 39.68 on 2 and 12 DF,  p-value: 5.136e-06
```

Calculando el factor de inflación de la varianza

```
vif(reg4)
```

```
##    medio temp_med
## 8.029523 8.029523
```

Es importante notar que existe multicolinealidad ya que, al analizar las pruebas t de las variables explicativas vemos que no son significativas pero al revisar la prueba F en conjunto vemos que se rechaza $H_0 = 0$, es decir las betas no son cero en conjunto, y como tenemos una R cuadrada elevada podemos concluir que existe multicolinealidad, ademas de que el factor de inflación de la varianza tambien es mayor a 5 y cercano a 10. Es algo que esperabamos que ya la variable temp_med es una combinación lineal de temperatura y medio, por lo tanto tenemos que variable dejaremos si la combinación lineal o las variables por sí solas.

Comparando la R cuadrada de las regresiones las varaibles medio y temp_med respectivamente obtenemos 0.8464 y 0.8344, es decir muy similares, por lo tanto como sabemos que temp_med es una combinación lineal de medio, nos podemos quedar solo con alguna de estas dos variables, como la prueba de hipoetsis es que la temepratura y el medio afectan el conteo de las bacterias hacemos la regresión.

```
reg6=lm(logconteo~temperatura+medio, data=datos)
summary(reg6)
```

```
##
## Call:
## lm(formula = logconteo ~ temperatura + medio, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02711 -0.40755  0.05737  0.58505  0.74021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.402375   0.699897   0.575    0.576
## temperatura 0.008753   0.022909   0.382    0.709
## medio       1.216399   0.152734   7.964 3.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.6714 on 12 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8229
## F-statistic: 33.53 on 2 and 12 DF,  p-value: 1.222e-05
```

Como podemos observar la temperatura sigue siendo no significativa, ademas observamos que el intercepto tampoco es significativo ademas de que es cercano a 0, pero se rechaza $H_0 = 0$ para la prueba conjunta F.

Viendo el intervalo de confinaza para los estimadores para ver si es viable eliminar el intercepto

```
conf=confint(reg6, level=.90)
conf
```

```
##                      5 %       95 %
## (Intercept) -0.84504364 1.64979311
## temperatura -0.03207842 0.04958407
## medio        0.94418419 1.48861438
```

Notamos que en efecto, el intercepto se encuentra cerca del 0 y al ser no significativo decidimos retirarlo de la regresión.

```
reg7=lm(logconteo~temperatura+medio-1, data=datos)
summary(reg7)
```

```
##
## Call:
## lm(formula = logconteo ~ temperatura + medio - 1, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97479 -0.45756  0.01429  0.54958  0.83194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## temperatura  0.02017    0.01113   1.812   0.0931 .
## medio        1.23950    0.14351   8.637 9.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6538 on 13 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9646
## F-statistic: 205.2 on 2 and 13 DF,  p-value: 1.467e-10
```

Notamos que tenemos una R cuadrada muy cercana a 1, por lo tanto nos quedaremos con este modelo, pero rechazamos la hipotesis que se plantea al inicio ya que como vimos la temperatura no es significativa, solo el medio es significativo en la regresión.

Verificación de Supuestos

Autocorrelación

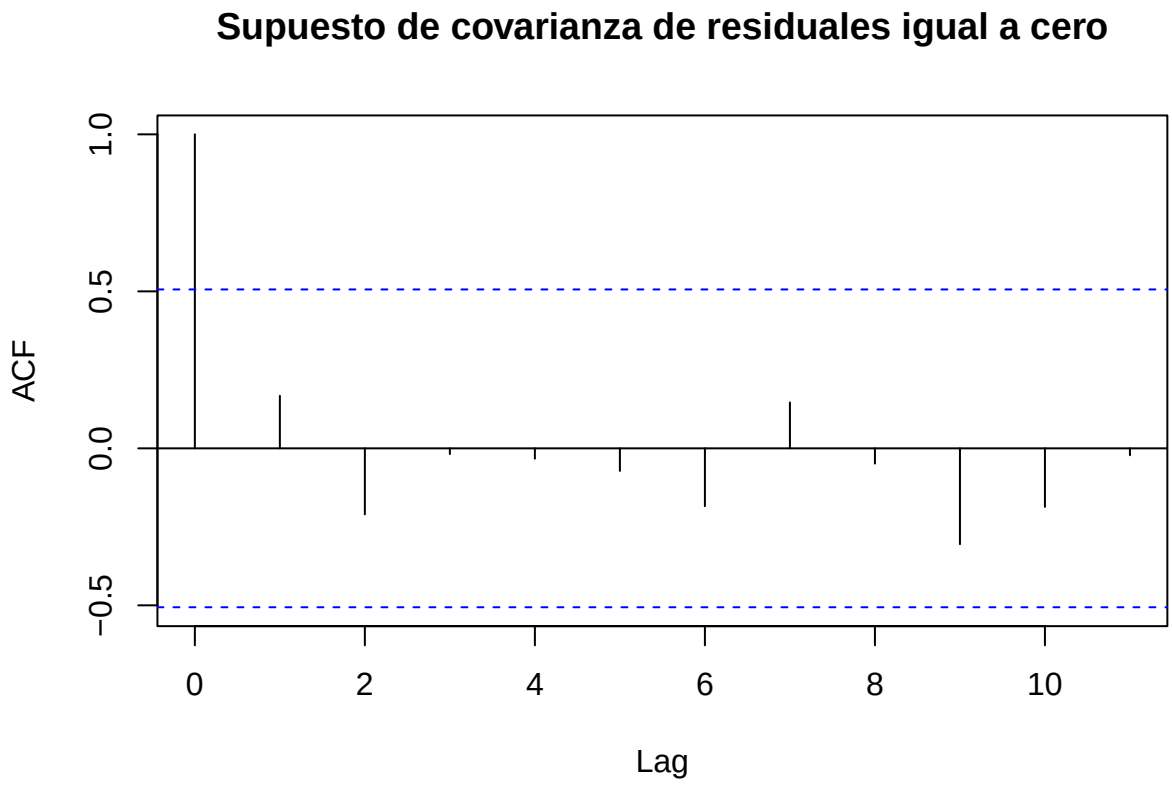
Primero haremos una prueba de hipotesis Durbin-Watson

```
dwtest(reg7)
```

```
##
##  Durbin-Watson test
##
## data:  reg7
## DW = 1.5912, p-value = 0.1981
## alternative hypothesis: true autocorrelation is greater than 0
```

Observamos que el valor DW es cercano a 2, con una varinaza significativa pero todavía aceptable como que no hay autocorrelación, haremos una correlograma para ver graficamente si es que no existe correlación.

```
acf(reg7$residuals, main="Supuesto de covarianza de residuales igual a cero")
```



Se cumple el supuesto de que no existe autocorrelación.

Esperanza de los residuales igual a cero

```
bgtest(reg7)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  reg7
## LM test = 0.43196, df = 1, p-value = 0.511
```

Por lo tanto no se rchaza la hipotesis nula, es decir la esperanza de los residuales es cero con un 95% de efectividad.

Homocedasticidad

```
bptest(reg7)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg7
## BP = 0.14632, df = 1, p-value = 0.7021
```

Por lo tanto no se rechaza la hipotesis nula, es decir la varianza de los errores es constante.

Normalidad de los residuales

Haremos una prueba Shapiro-Wilks y un qq-plt para la prueba gráfica

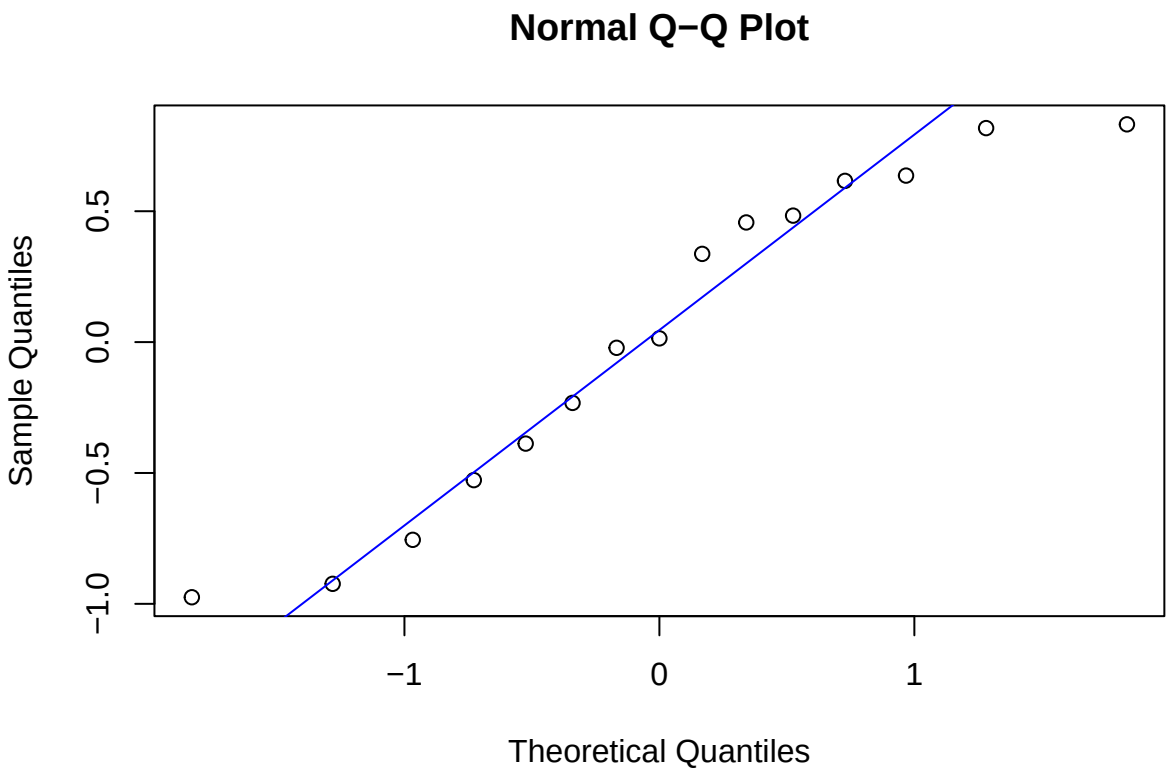
```
shapiro.test(reg7$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg7$residuals
## W = 0.92235, p-value = 0.2092
```

No se rechaza H_0= Se distribuye normal, es decir se acepta la hipotesis nula.

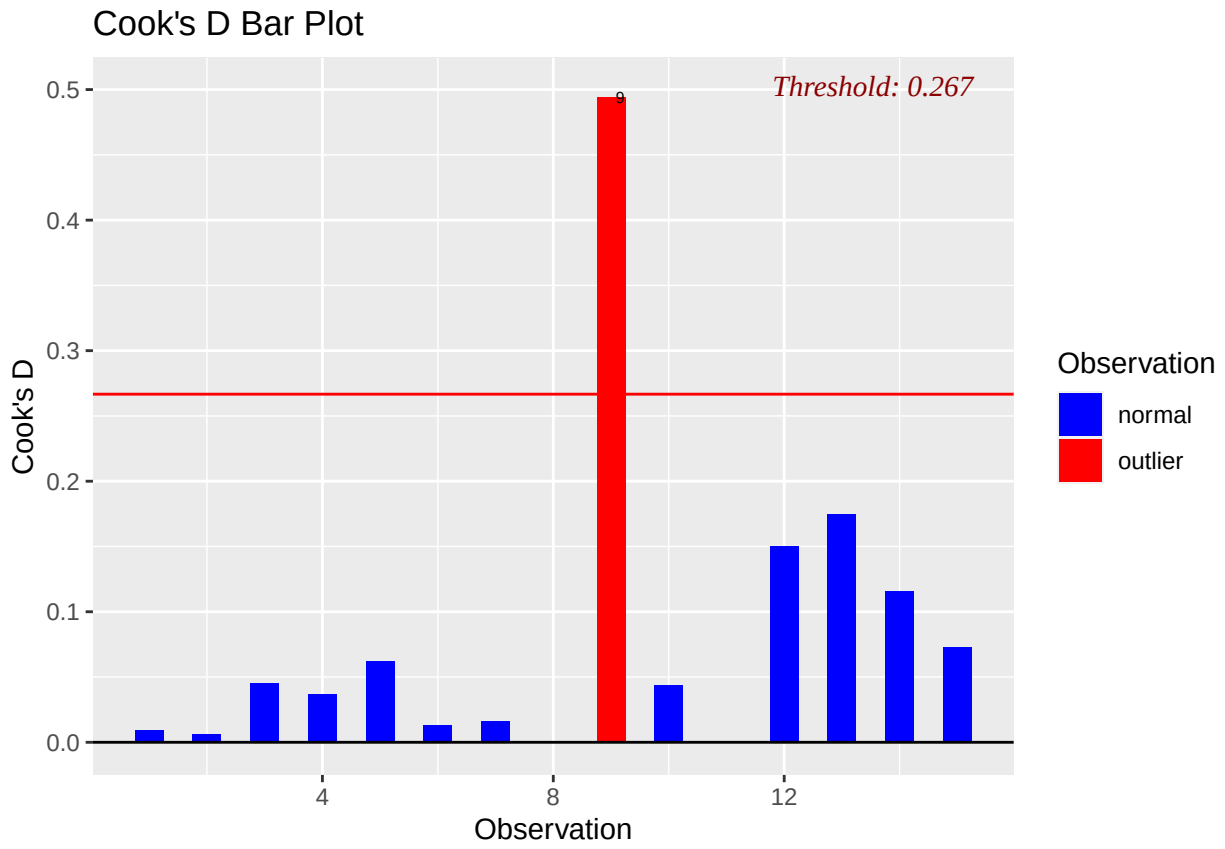
Haciendo un qq-plot

```
qqnorm(reg7$residuals)
qqline(reg7$residuals, col ="blue")
```



Notamo que existen 3 datos que no se ajustan bien a la recta, que probablemente sean outliers, asi que haremos una gráfica con la distancia de Cooks para identificar outliers

```
ols_plot_cooksd_bar(reg7)
```



Observamos que solo tenemos un outlier en la observación número 9, donde la variable temperatura alcanza su máximo con 45 grados, ya que al tener un modelo bueno dada la R cuadrada obtenida, decidimos dejar el outlier ya que no nos generó problemas al modelar.

```
summary(reg7)
```

```
##
## Call:
## lm(formula = logconteo ~ temperatura + medio - 1, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97479 -0.45756  0.01429  0.54958  0.83194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## temperatura  0.02017    0.01113   1.812   0.0931 .
## medio        1.23950    0.14351   8.637 9.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6538 on 13 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9646
## F-statistic: 205.2 on 2 and 13 DF,  p-value: 1.467e-10
```

Conclusiones

Podemos concluir que la hipotesis planteada de que la temperatura y el medio afectan el conteo de bacterias no es del todo cierto, ya que al hacer la regresión vimos que la variable temperatura no era significante, pero al final decidimos dejarla ya que el proposito era ver si se cumplia la hipotesis, teniendo como resultado que el medio es la variable mas influyente en la regresión y dado el intercepto estimado es mejor quitarlo ya que podemos aumentar el valor de la R cuadrada. Se concluye que existe una relación lineal positiva entre las variables medio y temperatura, siendo el medio es mas influyente ya que por cada unidad que aumenta en el medio, el conteo de bacterias aumenta 1.23 veces.