Práctica 4: Regresión lineal múltiple

Módulo de Modelos Lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

1. Para el banco de datos `Auto` de la librería `ISLR`, considera ahora un modelo de regresión lineal múltiple sobre `mpg` empleando como covariables, sin interacción, el resto de variables NUMÉRICAS del banco de datos (por tanto excluye la variable `origin` que sería una variable categórica). Interpreta los resultados obtenidos: ¿Qué variables tienen un efecto significativo sobre `mpg`? ¿De qué forma influyen? Respecto al efecto de `horsepower`, que ya fue estudiado en la práctica anterior ¿Cómo cambia la interpretación del efecto de dicha variable del análisis anterior a éste? ¿Cómo ha cambiado la variabilidad de su efecto?

```
# Cargamos el banco de datos
data(Auto, package = "ISLR")

# Modelo de regresión lineal simple para horsepower
modelo.old <- lm(mpg ~ horsepower, data = Auto)
summary(modelo.old)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
# Modelo de regresión lineal múltiple
modelo <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year,
    data = Auto)
summary(modelo)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6927 -2.3864 -0.0801  2.0291 14.3607
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
## cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
## displacement  7.678e-03  7.358e-03   1.044  0.29733
## horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
## weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
## acceleration  8.527e-02  1.020e-01   0.836  0.40383
## year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.435 on 385 degrees of freedom
## Multiple R-squared:  0.8093, Adjusted R-squared:  0.8063
## F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

```
# Lás unicas variables que tienen ahora efecto significativo son el peso y el año.

# El peso reduce las millas recorridas con un galón de gasolina.  El año, por el
# contrario, aumenta dicho número.

# La potencia del coche, una vez conocemos su peso y el año de construcción se hace
# irrelevante.

# La variabilidad del efecto de la potencia aumenta en el modelo de regresión lineal
# múltiple, posiblemente como efecto de la colinealidad con el resto de variables.  En
# parte la pérdida de significatividad se puede deber al aumento de la variabilidad
# del efecto de esta variable.
```

2. Volviendo al modelo que hayas ajustado en la tarea anterior, considera ahora ese mismo modelo eliminando aquellas variables que no tuvieran efecto significativo sobre `mpg`.

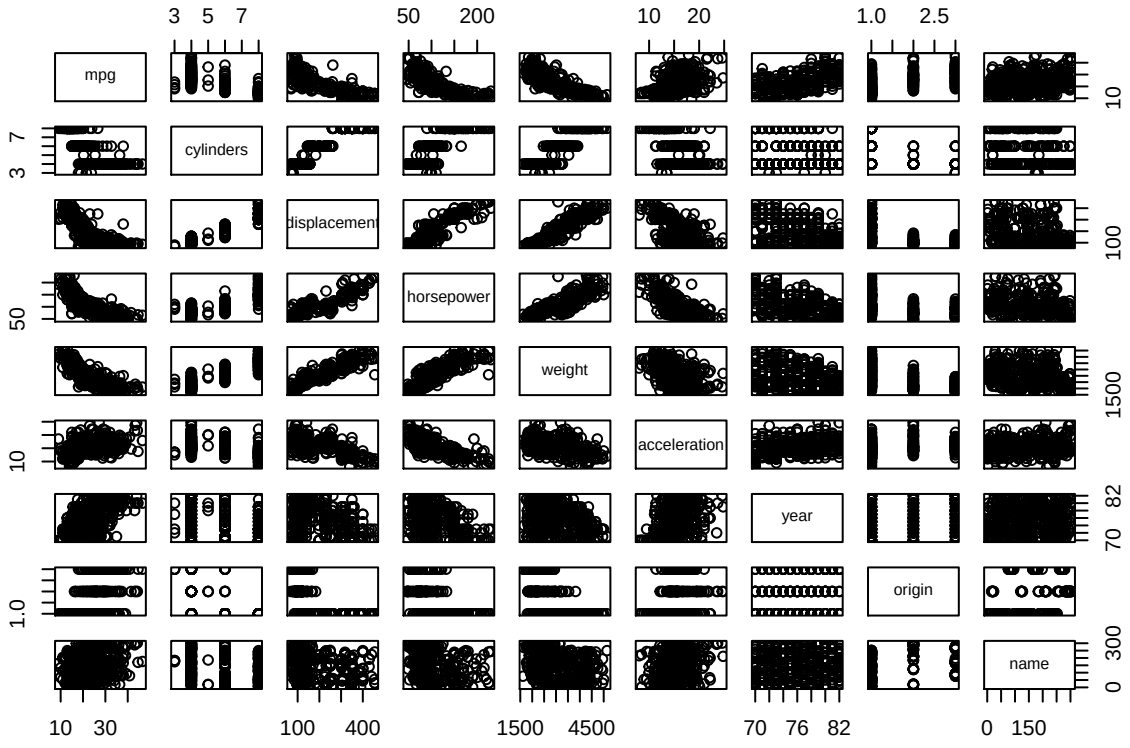
- ¿Cómo varía la estimación y variabilidad de dichos efectos al eliminar el resto de variables del modelo?
- ¿A qué crees que se debe la distinta disminución en la varianza de las estimadores en este nuevo modelo?
- ¿Consideras aconsejable considerar en el modelo la interacción de las variables que hubieras incluido en el modelo anterior? Interpreta el efecto de la interacción que hayas incluido sobre `mpg`.

```
# Modelo con sólo covariables significativas.
modelo2 <- lm(mpg ~ weight + year, data = Auto)
summary(modelo2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8505 -2.3014 -0.1167  2.0367 14.3555
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.435e+01  4.007e+00  -3.581 0.000386 ***
## weight      -6.632e-03  2.146e-04 -30.911  < 2e-16 ***
## year         7.573e-01  4.947e-02  15.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.427 on 389 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8072
## F-statistic: 819.5 on 2 and 389 DF,  p-value: < 2.2e-16
```

```
# La variabilidad de las covariables ha disminuido respecto el modelo anterior,
# particularmente para el peso, posiblemente porque este factor se correlacione con
# bastantes más variables que el año. En consecuencia sus p-valores se vuelven más
# significativos (fíjate en los estadísticos t), particularmente para el peso.

plot(Auto)
```




```
# Aquí podemos ver que el peso se correlaciona con bastantes más factores que el año.

# Modelo con interacción entre peso y año
modelo3 <- lm(mpg ~ weight * year, data = Auto)
summary(modelo3)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ weight * year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339, Adjusted R-squared:  0.8326
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
# Sí que resulta aconsejable incluir la interacción ya que tiene un efecto
# significativo sobre la variable respuesta.

summary(cbind(Auto$weight, Auto$year))
```

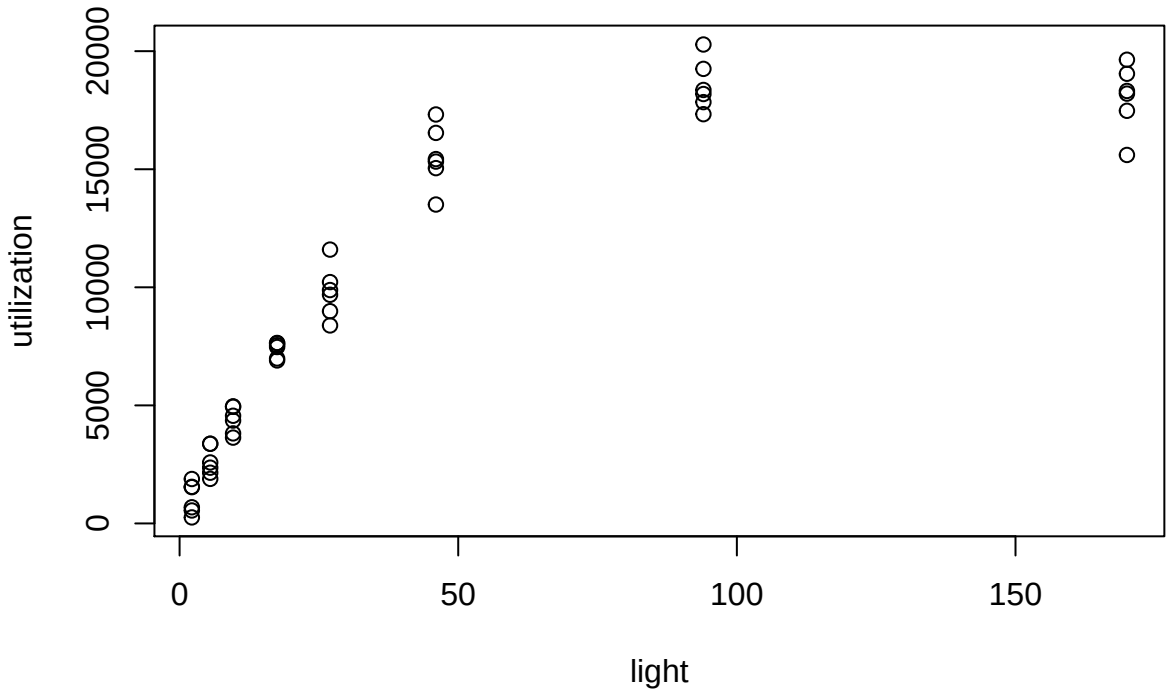
```
##        V1             V2
##  Min.   :1613   Min.   :70.00
##  1st Qu.:2225   1st Qu.:73.00
##  Median :2804   Median :76.00
##  Mean   :2978   Mean   :75.98
##  3rd Qu.:3615   3rd Qu.:79.00
##  Max.   :5140   Max.   :82.00
```

```
# Efecto del peso: E(mpg)= (2.7*10^{-2}-4.58*10^{-4}*year)*weight+ ...

# year varía entre 70 y 82 -> El efecto del peso sobre mpg pasa de ser -0.0045 al
# principio del periodo (más peso menos millas) a -0.0100 (más peso muchas menos
# millas), por tanto el peso es más importante al final del periodo.

# Efecto del año: E(mpg)= (2.04-4.58*10^{-4}*weight)*year+ ...

# weight varía entre 1613 y 5140 -> El efecto del año sobre mpg pasa de ser de 1.3
# para el coche más ligero (ha mejorado su consumo) a -0.31 para el más pesado (ha
# empeorado su consumo o posiblemente no ha mejorado).
```

3. El archivo `Nitrite.Rdata` contiene datos sobre el consumo de nitritos de distintas plantas de alubias en función de la intensidad de la luz.

- Representa la nube de puntos correspondiente a este banco de datos, ignora la variable día para el análisis estadístico.
- Ajusta la relación polinomial que consideres más adecuada para representar la relación entre el uso de nitritos e intensidad de la luz.
- Representa gráficamente la relación polinomial que te haya parecido oportuna, así como la relación correspondiente al resto de polinomios de orden menor que hayas ajustado para observar las diferencias que se producen entre los distintos ajustes.
- Para el modelo que consideres más adecuado, representa la relación ajustada entre nitritos e intensidad de luz, así como sus intervalos de confianza y predicción al 95%.

```r
load("../Datos/Nitrite.Rdata")

# Representación de la relación entre variables
with(Nitrite, plot(light, utilization))
```




```r
# Ajuste de modelos polinomiales de distintos grados.  Tomamos polinomios ortogonales
# ya que no nos interesa demasiado la interpretación de sus términos y presentan menor
# colinealidad.
Nit1 <- lm(utilization ~ poly(light, 1), data = Nitrite)
summary(Nit1)
```

```
##
## Call:
## lm(formula = utilization ~ poly(light, 1), data = Nitrite)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6462.1 -3230.9  -432.2  3020.0  7698.2
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9668.5      539.2   17.93  < 2e-16 ***
## poly(light, 1)  37878.2     3735.9   10.14 2.62e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3736 on 46 degrees of freedom
## Multiple R-squared:  0.6909, Adjusted R-squared:  0.6841
## F-statistic: 102.8 on 1 and 46 DF,  p-value: 2.618e-13
```

```
Nit2 <- lm(utilization ~ poly(light, 2), data = Nitrite)
summary(Nit2)
```

```
##
## Call:
## lm(formula = utilization ~ poly(light, 2), data = Nitrite)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3019.2  -808.7    58.2   750.7  3593.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9668.5      201.2   48.05   <2e-16 ***
## poly(light, 2)1  37878.2     1394.1   27.17   <2e-16 ***
## poly(light, 2)2 -23549.0     1394.1  -16.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1394 on 45 degrees of freedom
## Multiple R-squared:  0.9579, Adjusted R-squared:  0.956
## F-statistic: 511.8 on 2 and 45 DF,  p-value: < 2.2e-16
```

```
Nit3 <- lm(utilization ~ poly(light, 3), data = Nitrite)
summary(Nit3)
```

```
##
## Call:
## lm(formula = utilization ~ poly(light, 3), data = Nitrite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2428.70  -517.23    66.61   557.02  2392.15
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9668.5      140.5  68.808  < 2e-16 ***
## poly(light, 3)1  37878.2      973.5  38.909  < 2e-16 ***
## poly(light, 3)2 -23549.0      973.5 -24.190  < 2e-16 ***
## poly(light, 3)3   6764.2      973.5   6.948 1.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 973.5 on 44 degrees of freedom
## Multiple R-squared:  0.9799, Adjusted R-squared:  0.9786
## F-statistic: 715.8 on 3 and 44 DF,  p-value: < 2.2e-16
```

```
Nit4 <- lm(utilization ~ poly(light, 4), data = Nitrite)
summary(Nit4)
```

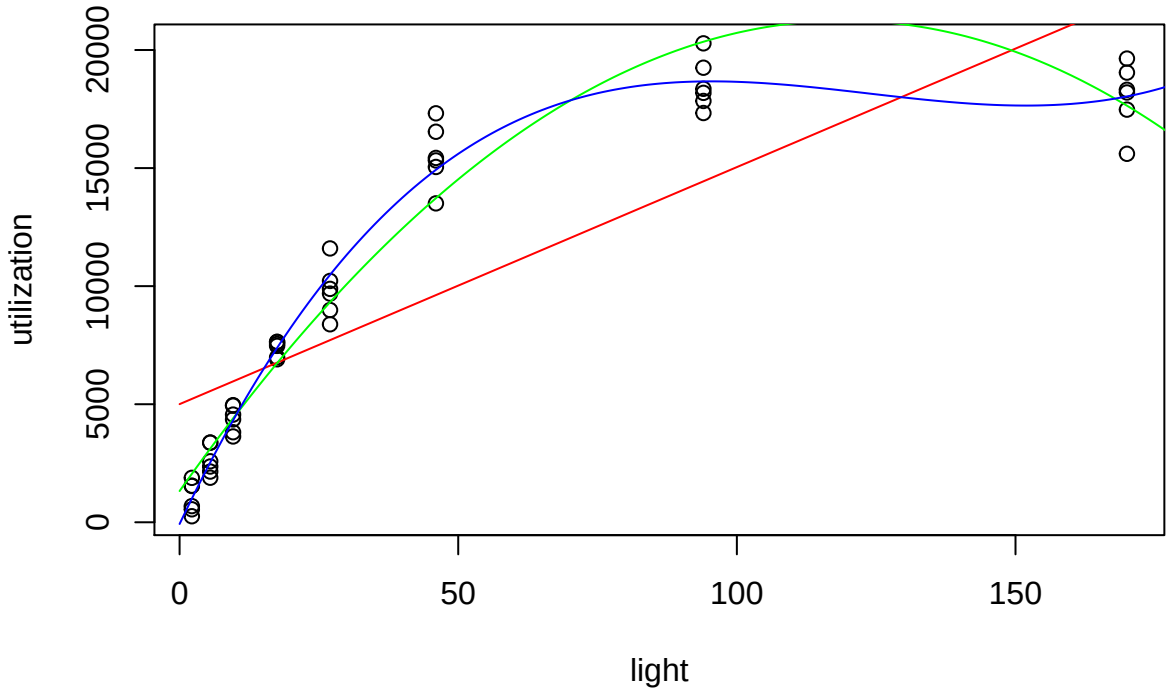
```
##
## Call:
```

```
## lm(formula = utilization ~ poly(light, 4), data = Nitrite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2440.52  -537.95    -7.39   566.83  1971.91
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9668.5      137.9  70.130  < 2e-16 ***
## poly(light, 4)1  37878.2      955.2  39.657  < 2e-16 ***
## poly(light, 4)2 -23549.0      955.2 -24.655  < 2e-16 ***
## poly(light, 4)3   6764.2      955.2   7.082  9.8e-09 ***
## poly(light, 4)4   1571.5      955.2   1.645    0.107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 955.2 on 43 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9794
## F-statistic: 558.3 on 4 and 43 DF,  p-value: < 2.2e-16
```

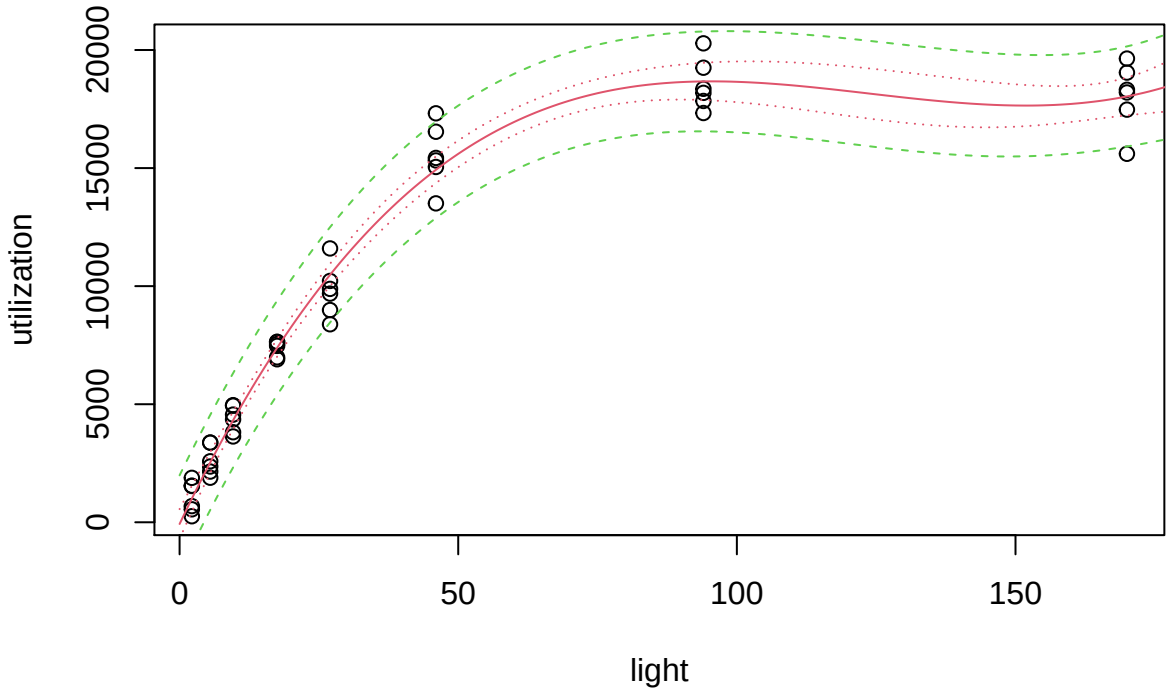
```
Nit5 <- lm(utilization ~ poly(light, 5), data = Nitrite)
summary(Nit5)
```

```
##
## Call:
## lm(formula = utilization ~ poly(light, 5), data = Nitrite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2441.32  -487.67   -97.99   500.64  1808.44
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9668.5      135.4  71.408  < 2e-16 ***
## poly(light, 5)1  37878.2      938.1  40.379  < 2e-16 ***
## poly(light, 5)2 -23549.0      938.1 -25.104  < 2e-16 ***
## poly(light, 5)3   6764.2      938.1   7.211 7.27e-09 ***
## poly(light, 5)4   1571.5      938.1   1.675    0.101
## poly(light, 5)5   1507.2      938.1   1.607    0.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 938.1 on 42 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9801
## F-statistic: 463.6 on 5 and 42 DF,  p-value: < 2.2e-16
```

```
# Parece que un polinomio de grado 3 podría ser adecuado

# Representación de los distintos ajustes
plot(utilization ~ light, data = Nitrite)
x <- 0:180
lines(x, predict(Nit1, newdata = data.frame(light = x)), col = "red")
lines(x, predict(Nit2, newdata = data.frame(light = x)), col = "green")
lines(x, predict(Nit3, newdata = data.frame(light = x)), col = "blue")
```




```
# Representación del ajuste cúbico con ICs e intervalos de predicción
plot(utilization ~ light, data = Nitrite)
lines(x, predict(Nit3, newdata = data.frame(light = x)), col = 2)
lines(x, predict(Nit3, newdata = data.frame(light = x), interval = "prediction")[, 2],
    col = 3, lty = 2)
lines(x, predict(Nit3, newdata = data.frame(light = x), interval = "prediction")[, 3],
    col = 3, lty = 2)
lines(x, predict(Nit3, newdata = data.frame(light = x), interval = "confidence")[, 2],
    col = 2, lty = 3)
lines(x, predict(Nit3, newdata = data.frame(light = x), interval = "confidence")[, 3],
    col = 2, lty = 3)
```



4. Plantéate la veracidad o falsedad de las siguientes afirmaciones:

- La suma de cuadrados residual siempre disminuirá al incluir nuevas variables en un modelo de regresión lineal múltiple.

```
# Efectivamente, al disponer de una variable más, el modelo de regresión obtendrá una
# variabilidad residual inferior que la de cualquier modelo al que le falte alguna de
# las variables que éste contiene. Al fin y al cabo el modelo sin la covariable
# oportuno estará restringido a que algunas de sus coeficientes sea 0 mientras que el
# modelo que contiene todas las covariables no presenta restricción alguna.
```

- La variabilidad de los coeficientes de un modelo de regresión siempre disminuirá al incluir nuevas variables en el modelo.

```
# No, la colinealidad puede perfectamente aumentar dicha variabilidad.
```

- La variabilidad de los coeficientes de un modelo de regresión siempre aumentará al incluir nuevas variables en el modelo.

```
# No necesariamente, observa los resultados del ejercicio 3. Si las covariables son
# ortogonales la variabilidad de los coeficientes no tiene porqué aumentar sino que en
# general disminuirá.
```

- Si la correlación de la variable respuesta con cierta covariable es positiva, el correspondiente coeficiente en un modelo de regresión lineal múltiple será también positivo.

```
# En regresión lineal simple vimos que este resultado era cierto, pero no en regresión
# múltiple ya que en este caso el coeficiente de cierta covariable dependerá a su vez
# de otras covariables, por lo que el resultado mencionado no tiene porqué cumplirse.
```