6. Métodos Alternativos a la Regresión

bas, magí

2023-05-22

Ejercicios de libro de Faraway

1. (Ejercicio 7 cap. 7 pág. 110); Use the happy dataset with happy as the response and the other variables as predictors. Discuss possible rescalings of the variables with the aim of helping the interpretation of the model fit

```
library(faraway)

data(happy)

head(happy)
```

	happy <dbl>	money <dbl>	sex <dbl>	love <dbl>	work <dbl>
1	10	36	0	3	4
2	8	47	1	3	1
3	8	53	0	3	5
4	8	35	1	3	3
5	4	88	1	1	2
6	9	175	1	3	4

6 rows

```
?happy
```

love, work and happiness

Description

Data were collected from 39 students in a University of Chicago MBA class

Format

A data frame with 39 observations on the following 5 variables.

happy
Happiness on a 10 point scale where 10 is most happy

money
family income in thousands of dollars

sex

1 = satisfactory sexual activity, 0 = not

love
1 = lonely, 2 = secure relationships, 3 = deep feeling of belonging and caring

work
5 point scale where 1 = no job, 3 = OK job, 5 = great job

Source

George and McCulloch (1993) “Variable Selection via Gibbs Sampling” JASA, 88, 881-889

```
# Creamos el modelo
happy_model= lm(happy ~ ., data=happy)

summary(happy_model)
```

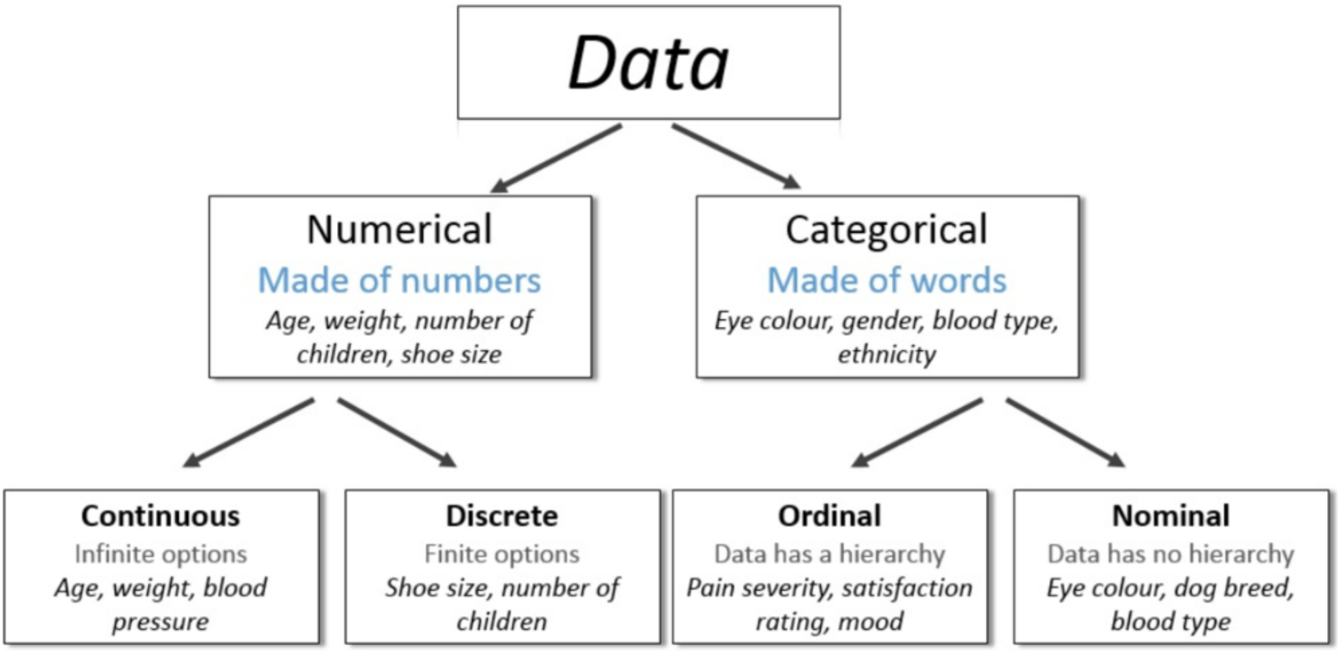
```
## 
## Call:
## lm(formula = happy ~ ., data = happy)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7186 -0.5779 -0.1172  0.6340  2.0651 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.072081   0.852543  -0.085   0.9331    
## money        0.009578   0.005213   1.837   0.0749 .  
## sex         -0.149008   0.418525  -0.356   0.7240    
## love         1.919279   0.295451   6.496 1.97e-07 ***
## work         0.476079   0.199389   2.388   0.0227 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.058 on 34 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.6761 
## F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

Un posible rescalado de variables para mejorar la interpretación del modelo podría ser normalizar las variables continuas para que tengan una media de cero y una desviación estándar de uno.

```
# Averiguamos el tipo de variables del dataset para el resclado
str(happy)
```

```
## 'data.frame':    39 obs. of  5 variables:
##  $ happy: num  10 8 8 8 4 9 8 6 5 4 ...
##  $ money: num  36 47 53 35 88 175 175 45 35 55 ...
##  $ sex  : num  0 1 0 1 1 1 1 0 1 1 ...
##  $ love : num  3 3 3 3 1 3 3 2 2 1 ...
##  $ work : num  4 1 5 3 2 4 4 3 2 4 ...
```

Recordar:


Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

```
happy$money_norm <- scale(happy$money)
happy_formula <- happy ~ .
happy_model2 <- lm(happy_formula, data = happy)
summary(happy_model2)
```

```
##
## Call:
## lm(formula = happy_formula, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.5779 -0.1172  0.6340  2.0651
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.072081   0.852543  -0.085   0.9331
## money        0.009578   0.005213   1.837   0.0749 .
## sex         -0.149008   0.418525  -0.356   0.7240
## love         1.919279   0.295451   6.496 1.97e-07 ***
## work         0.476079   0.199389   2.388   0.0227 *
## money_norm         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 34 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.6761
## F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

2. (Ejercicio 1 cap. 8 pág. 130); Researchers at National Institutes of Standards and Technology (NIST) collected pipeline data on ultrasonic measurements of the depth of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. It turns out that this batch effect is not significant and so can be ignored in the analysis that follows. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive. We want to develop a regression equation for correcting the in-field measurements.

(a) Fit a regression model Lab ~ Field. Check for non-constant variance

```
library(faraway)
data(pipeline)
names(pipeline)
```

```
## [1] "Field" "Lab"   "Batch"
```

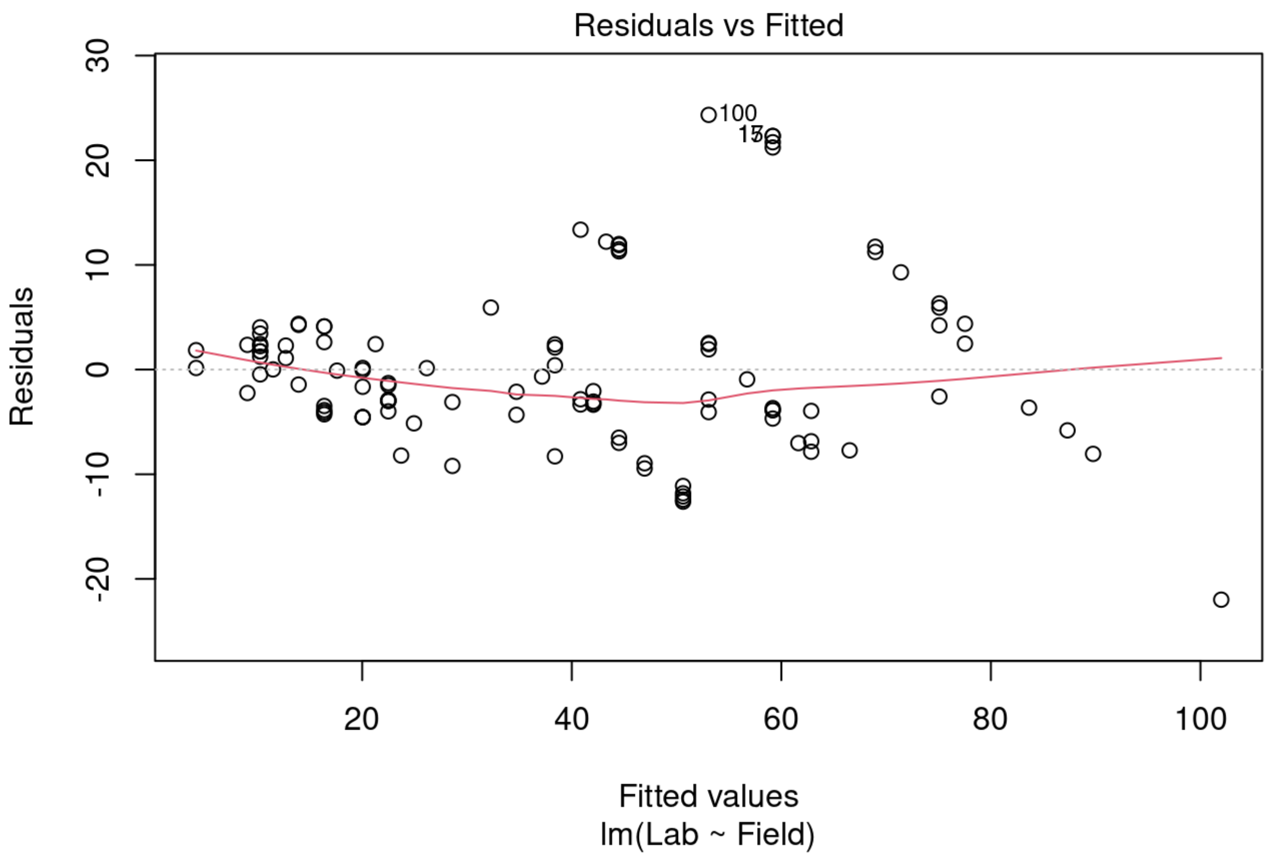
```
pipelinie_model= lm(Lab ~ Field, pipeline)
summary(pipelinie_model)
```

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

Una “non-constant variable” se refiere a una variable cuya varianza no es constante en todos los niveles de las variables predictoras en un modelo de regresión. En otras palabras, la variabilidad de los errores o residuos del modelo no es la misma en todas las condiciones o valores de las variables independientes.

Cuando la varianza no es constante, puede haber una tendencia de los errores a aumentar o disminuir a medida que cambian los valores de las variables predictoras. Esto puede manifestarse en el gráfico de “residuales vs. ajustados” como un patrón en forma de embudo o una dispersión irregular de los residuos alrededor de la línea horizontal en y = 0.

```
# Check for non-constant variable
plot(pipelinie_model,1)
```



En nuestro grafico no detectamos ningun indicio de problemas con una varianza no contante de la variable ‘Field’ ya que la línea roja sigue una linea horizontal suficientemente constante.

(b) We wish to use weights to account for the non-constant variance. Here we split the range of Field into 12 groups of size nine (except for the last group which has only eight values). Within each group, we compute the variance of Lab as varlab and the mean of Field as meanfield. Supposing pipeline is the name of your data frame, the following R code will make the needed computations:

```
i <- order(pipeline$Field)

npipe <- pipeline[i,]

ff <- gl(12,9)[-108]

meanfield <- unlist(lapply(split(npipe$Field,ff),mean))

varlab <- unlist(lapply(split(npipe$Lab,ff),var))
```

Suppose we guess that the the variance in the response is linked to the predictor in the following way:

$$var(Lab) = a_0 Field^{a_1}$$

Regress log(varlab) on log(meanfield) to estimate a0 and a1. (You might choose to remove the last point.) Use this to determine appropriate weights in a WLS fit of Lab on Field. Show the regression summary.

```
# Crear un data frame con los minimos quadrados ajustados
wls_data= data.frame(log_meanfield = log(meanfield), log_varlab = log(varlab))

# Ajustar un modelo de regresión de log(varlab) en log(meanfield)
reg_model= lm(log_varlab ~ log_meanfield, data = wls_data)

# Imprimir el resumen de la regresión
summary(reg_model)
```

```
##
## Call:
## lm(formula = log_varlab ~ log_meanfield, data = wls_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2038 -0.6729  0.1656  0.7205  1.1891
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.3538     1.5715  -0.225   0.8264
## log_meanfield   1.1244     0.4617   2.435   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 10 degrees of freedom
## Multiple R-squared:  0.3723, Adjusted R-squared:  0.3095
## F-statistic: 5.931 on 1 and 10 DF,  p-value: 0.03513
```

```
# Calcular la varianza predicha basada en la relación estimada
varianza_predicha= exp(predict(reg_model, newdata = data.frame(log_meanfield = log(pi
peline$Field))))

# Calcular los pesos como el recíproco de la varianza predicha
pesos= 1 / varianza_predicha

# Ajustar el modelo WLS
modelo_wls= lm(Lab ~ Field, data = pipeline, weights = pesos)

# Imprimir el resumen de la regresión del modelo WLS
summary(modelo_wls)
```

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline, weights = pesos)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0826 -0.8102 -0.3189  0.6212  3.4429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.49436    0.90707  -1.647    0.102
## Field        1.20828    0.03488  34.637   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.169 on 105 degrees of freedom
## Multiple R-squared:  0.9195, Adjusted R-squared:  0.9188
## F-statistic:  1200 on 1 and 105 DF,  p-value: < 2.2e-16
```

(c) An alternative to weighting is transformation. Find transformations on Lab and/or Field so that in the transformed scale the relationship is approximately linear with constant variance. You may restrict your choice of transformation to square root, log and inverse

Podeos provar con: logaritmo y raiz quadrada

```
# Transformación de Lab y Field utilizando raíz cuadrada
pipeline$Lab_sqrt <- sqrt(pipeline$Lab)
pipeline$Field_sqrt <- sqrt(pipeline$Field)

# Ajustar un modelo de regresión lineal en la escala transformada
model_sqrt <- lm(Lab_sqrt ~ Field_sqrt, data = pipeline)

# Imprimir el resumen del modelo
summary(model_sqrt)
```

```
##
## Call:
## lm(formula = Lab_sqrt ~ Field_sqrt, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1570 -0.4125 -0.1209  0.3098  1.5481
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.36773    0.18815  -1.954   0.0533 .
## Field_sqrt   1.13553    0.03247  34.973   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5561 on 105 degrees of freedom
## Multiple R-squared:  0.9209, Adjusted R-squared:  0.9202
## F-statistic:  1223 on 1 and 105 DF,  p-value: < 2.2e-16
```

```
# Transformación de Lab y Field utilizando logaritmo
pipeline$Lab_log <- log(pipeline$Lab)
pipeline$Field_log <- log(pipeline$Field)

# Ajustar un modelo de regresión lineal en la escala transformada
model_log <- lm(Lab_log ~ Field_log, data = pipeline)

# Imprimir el resumen del modelo
summary(model_log)
```

```
##
## Call:
## lm(formula = Lab_log ~ Field_log, data = pipeline)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40212 -0.11853 -0.03092  0.13424  0.40209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06849    0.09305  -0.736    0.463
## Field_log    1.05483    0.02743  38.457   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1837 on 105 degrees of freedom
## Multiple R-squared:  0.9337, Adjusted R-squared:  0.9331
## F-statistic:  1479 on 1 and 105 DF,  p-value: < 2.2e-16
```

3. (Ejercicio 2 cap. 8 pág. 131); Using the divusa data, fit a regression model with divorce as the response and unemployed, femlab, marriage, birth and military as predictors.

```
data(divusa)
?divusa
```

Divorce in the USA 1920-1996

Description

Divorce rates in the USA from 1920-1996

Format

A data frame with 77 observations on the following 7 variables.

year
the year from 1920-1996

divorce
divorce per 1000 women aged 15 or more

unemployed
unemployment rate

femlab
percent female participation in labor force aged 16+

marriage
marriages per 1000 unmarried women aged 16+

birth
births per 1000 women aged 15-44

military
military personnel per 1000 population

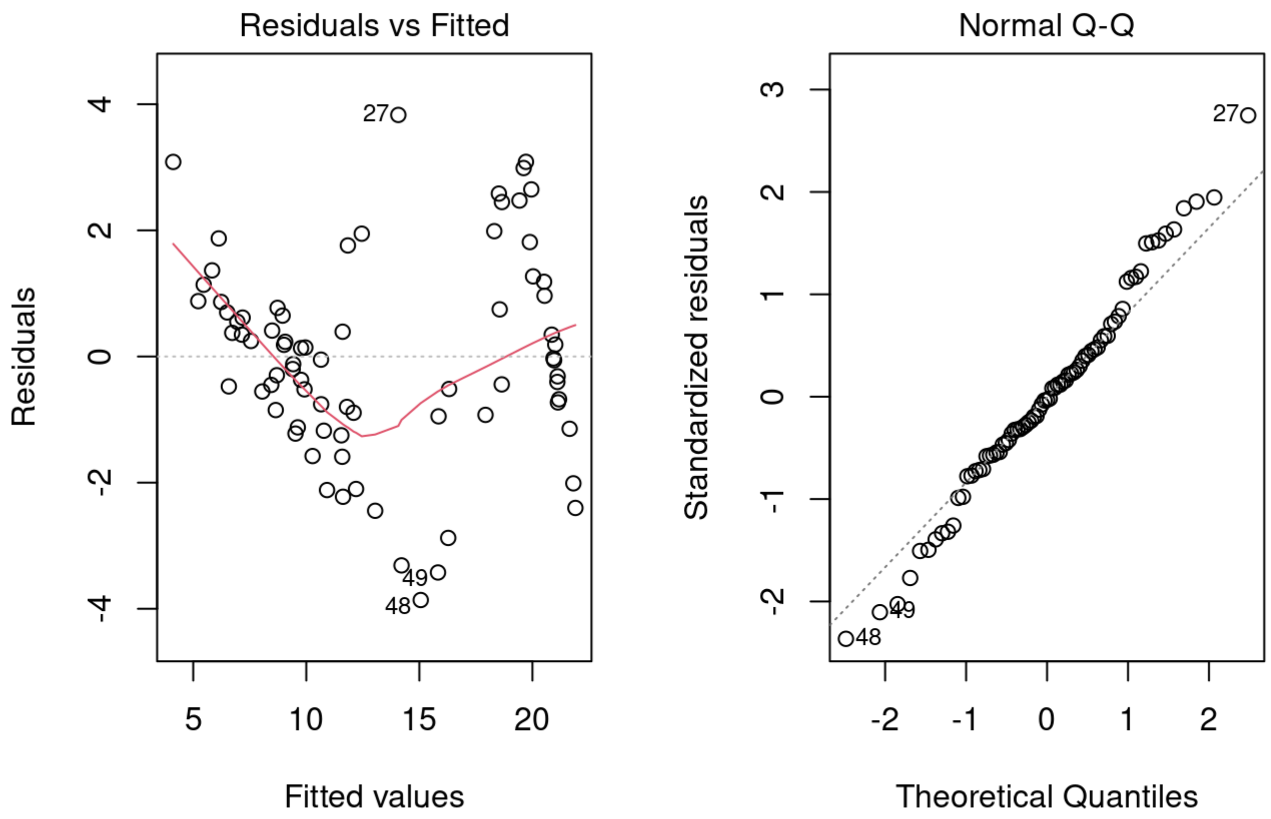
```
# Fit a regression model with divorce as the response and unemployed, femlab, marriag
e, birth and military as predictors
divusa_model= lm(divorce ~ unemployed + femlab + marriage + birth + military, divusa)
summary(divusa_model)
```

```
##
## Call:
## lm(formula = divorce ~ unemployed + femlab + marriage + birth +
##     military, data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8611 -0.8916 -0.0496  0.8650  3.8300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48784    3.39378   0.733   0.4659
## unemployed  -0.11125    0.05592  -1.989   0.0505 .
## femlab       0.38365    0.03059  12.543  < 2e-16 ***
## marriage     0.11867    0.02441   4.861 6.77e-06 ***
## birth       -0.12996    0.01560  -8.333 4.03e-12 ***
## military    -0.02673    0.01425  -1.876   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.65 on 71 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9152
## F-statistic: 165.1 on 5 and 71 DF,  p-value: < 2.2e-16
```

(a) Make two graphical checks for correlated errors. What do you conclude?

```
par(mfrow= c(1,2))

p1= plot(divusa_model, 1)
p2= plot(divusa_model, 2)
```



```
par(mfrow = c(1,1))
```

Con el plot1, podemos decir que seguramente hay un problema de linealidad con el modelo.

En el plot2, vemos como los residuos se ajustan a la línia diagonal, excepto en los extremos. Este gráfico nos sirve para avaluar la normalidad de residuos.

(b) Allow for serial correlation with an AR(1) model for the errors. (Hint: Use maximum likelihood to estimate the parameters in the GLS fit by gls(…,method="ML", …)). What is the estimated correlation and is it significant? Does the GLS model change which variables are found to be significant?

```
library(nlme)

# Metodo de mínimos cuadrados generalizados (GLS) con estimación de máxima verosimilitud
divusa_model_ar1 = gls(divorce ~ unemployed + femlab + marriage + birth + military, data = divusa, method = "ML")

summary(divusa_model_ar1)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: divorce ~ unemployed + femlab + marriage + birth + military
##   Data: divusa
##        AIC      BIC    logLik
##   303.4288 319.8355 -144.7144
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept)  2.4878446  3.393779  0.733060  0.4659
## unemployed  -0.1112520  0.055925 -1.989319  0.0505
## femlab       0.3836493  0.030587 12.542988  0.0000
## marriage     0.1186743  0.024414  4.860852  0.0000
## birth       -0.1299592  0.015595 -8.333384  0.0000
## military    -0.0267340  0.014247 -1.876433  0.0647
##
##  Correlation:
##            (Intr) unmply femlab marrig birth
## unemployed -0.745
## femlab     -0.928  0.652
## marriage   -0.687  0.373  0.572
## birth      -0.554  0.471  0.506 -0.171
## military    0.020  0.239 -0.063 -0.206  0.073
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.43627690 -0.56258793 -0.03128648  0.54578530  2.41666629
##
## Residual standard error: 1.584818
## Degrees of freedom: 77 total; 71 residual
```

Los coeficientes de correlación estimados son:

```
 (Intr) unmply femlab marrig birth
unemployed -0.745
femlab     -0.928  0.652
marriage   -0.687  0.373  0.572
birth      -0.554  0.471  0.506 -0.171
military    0.020  0.239 -0.063 -0.206  0.073
```

Se muestran los valores de correlación entre las distintas variables.

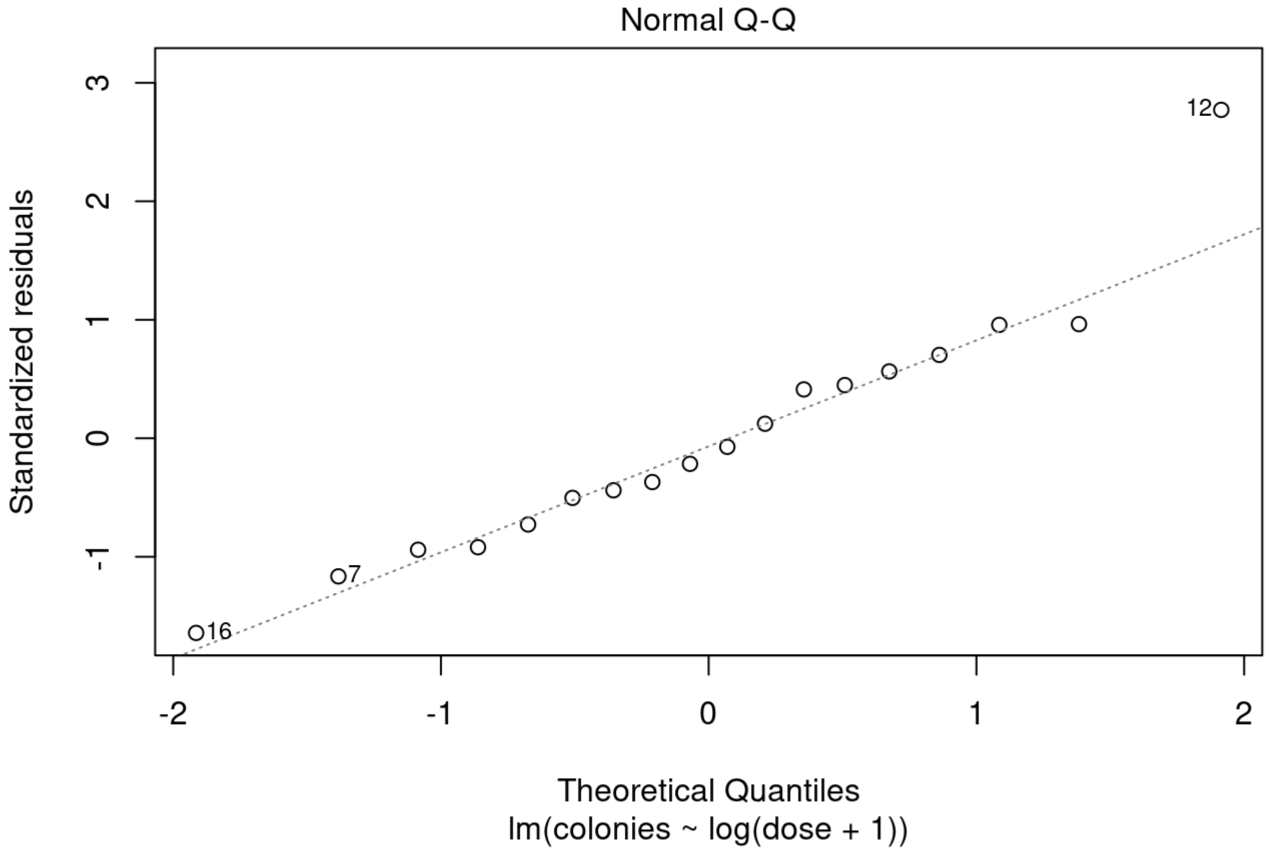
(c) Speculate why there might be correlation in the errors

4. (Ejercicio 3 cap. 8 pág. 131) For the salmonella dataset, fit a linear model with colonies as the response and log(dose+1) as the predictor. Check for lack of fit.

```
data(salmonella)
salmonella_model= lm(colonies ~ log(dose+1), salmonella)
summary(salmonella_model)
```

```
##
## Call:
## lm(formula = colonies ~ log(dose + 1), data = salmonella)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.376  -6.882  -1.509   5.400  29.119
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     19.823      5.064   3.915  0.00123 **
## log(dose + 1)    2.396      1.128   2.125  0.04955 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.84 on 16 degrees of freedom
## Multiple R-squared:  0.2201, Adjusted R-squared:  0.1713
## F-statistic: 4.514 on 1 and 16 DF,  p-value: 0.04955
```

```
plot(salmonella_model, 2)
```



5. (∗) (Ejercicio 4 cap. 8 pág. 131) For the cars dataset, fit a linear model with distance as the response and speed as the predictor. Check for lack of fit.

```
data(cars)
cars_model= lm(dist ~ speed, cars)
summary(cars_model)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Para evaluar el desempeño del ajuste tenemos dos grupos:

- Datos replicados (múltiples observaciones con los mismos valores para todas las variables predictoras):
  - Test de bondad de ajuste chi-cuadrado de Pearson
  - Test de bondad de ajuste de deviance (analogo al test de falta de ajuste F en regresión lineal múltiple)

Cuando se rechaza el test, hay una falta de ajuste estadísticamente significativa. De lo contrario, no hay evidencia de falta de ajuste.

- Conjuntos de datos NO replicados (o solo unas pocas observaciones replicadas):
  - Test de bondad de ajuste Hosmer-Lemeshow (solo para modelos logśiticos)

```
# Calculamos los residuos del modelo
residuals= resid(cars_model)

# Los agrupamos en categorias
categorias= cut(residuals, breaks = 5)

# Realizamos el test
chi_square= chisq.test(table(categorias))

chi_square
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(categorias)
## X-squared = 23, df = 4, p-value = 0.0001266
```

H0: no hay diferencia significativa entre las frecuencias observadas y las frecuencias esperadas bajo el modelo

H1: existe una diferencia significativa entre estas frecuencias p-value < 0.05; Rechazamos H0, el modelo lineal no se ajusta adecuadamente a los datos ‘cars’ y existe falta de ajuste.