Lista de ejercicios 4 - Análisis de regresión

Ciclo nivelación 2025-2

Mg. Sc. J. Eduardo Gamboa U.

Una universidad desea comprender qué factores influyen en el puntaje obtenido por los estudiantes en el examen final de un curso introductorio.

Para ello, recolecta información de 600 estudiantes sobre tres aspectos:

- Puntaje: Resultado del examen final (0–100 puntos), es la variable que se desea predecir.
- Estudio: Número de horas de estudio semanal dedicadas al curso.
- Curso: Modalidad en que el estudiante llevó el curso, la cual puede ser online, presencial o híbrido.
- Sueno: Número promedio de horas de sueño diario.

Se sospecha que:

- El puntaje aumenta con las horas de estudio.
- La modalidad del curso puede generar diferencias en rendimiento.
- Las horas de estudio pueden diferir entre las modalidades del curso.
- El efecto del sueño no es estrictamente lineal, sino potencialmente curvo.

Ajustar diversos modelos de regresión lineal considerando:

- el efecto de interacción, o no, de modalidad $\times$ estudio sobre el puntaje del examen.
- el efecto lineal, cuadrático o cúbico del sueño sobre el puntaje del examen.

Comparar estos modelos, con énfasis en su desempeño predictivo.

Lectura de datos

```
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.4.3
```

```
datos <- read_excel('Lista4_datos.xlsx')
datos |> head()
```

```
# A tibble: 6 x 4
  Puntaje Estudio Curso      Sueno
    <dbl>   <dbl> <chr>      <dbl>
1    98.2    19.3 Presencial  8.06
2    98.2    21.9 Presencial  5.85
3   100      21.9 Hibrido     5.69
4    80.4    10.1 Hibrido     7.47
5    75.9    13.7 Presencial  6.46
6    83.3    12.6 Online      7.1
```

```
datos |> nrow()
```

```
[1] 600
```

```
datos |> colnames()
```

```
[1] "Puntaje" "Estudio" "Curso"   "Sueno"
```

División en training y testing

```
n <- datos |> nrow()
set.seed(2025)
id    <- sample(1:n, 480)
train <- datos[id,]
test  <- datos[-id,]
train |> head()
```

```
# A tibble: 6 x 4
  Puntaje Estudio Curso      Sueno
    <dbl>   <dbl> <chr>      <dbl>
1    79.4   12.6  Presencial  8.95
2    91.5   12.1  Hibrido     6.17
3    84.4    8.66 Online      7.63
4    87.7   17.3  Hibrido     5.63
5    73.1   10.1  Online      7.16
6    90.6   13.9  Hibrido     6.58
```

```
test |> head()
```

```
# A tibble: 6 x 4
  Puntaje Estudio Curso      Sueno
    <dbl>   <dbl> <chr>      <dbl>
1    72.5    7.25 Online      8.36
2    92.6   19.0  Online      6.41
3    94.4   18.5  Presencial  7.37
4   100     20.3  Presencial  7.02
5    85.0   12.5  Hibrido     4.96
6   100     21.3  Online      7.96
```

Modelo 1

Modelo con interacción entre estudio y curso, y efecto cúbico del sueño

```
mod1 <- lm(Puntaje ~ Estudio*Curso + poly(Sueno,3), data = train)
mod1 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio * Curso + poly(Sueno, 3), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9682  -3.5372   0.2116   3.7794  14.9195

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             64.44252    1.34380  47.955   <2e-16 ***
Estudio                  1.59071    0.10270  15.489   <2e-16 ***
```

```
CursoOnline              -3.85034    1.79021  -2.151   0.0320 *
CursoPresencial          -1.46084    1.66633  -0.877   0.3811
poly(Sueno, 3)1          13.84045    5.43793   2.545   0.0112 *
poly(Sueno, 3)2         -13.90583    5.42517  -2.563   0.0107 *
poly(Sueno, 3)3           2.20569    5.42312   0.407   0.6844
Estudio:CursoOnline       0.12811    0.13745   0.932   0.3518
Estudio:CursoPresencial   0.05024    0.12753   0.394   0.6938
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.412 on 471 degrees of freedom
Multiple R-squared:  0.6996,    Adjusted R-squared:  0.6945
F-statistic: 137.1 on 8 and 471 DF,  p-value: < 2.2e-16
```

```
mod1 |> car::vif()
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                    GVIF Df GVIF^(1/(2*Df))
Estudio         4.118303  1        2.029360
Curso          52.633806  2        2.693495
poly(Sueno, 3)  1.018713  3        1.003095
Estudio:Curso  70.982042  2        2.902600
```

```
rtrain1 <- data.frame(real = train$Puntaje,
                      pred = predict(mod1, train))
library(caret)
```

```
Warning: package 'caret' was built under R version 4.4.3

Warning: package 'ggplot2' was built under R version 4.4.3

Warning: package 'lattice' was built under R version 4.4.3
```

```
postResample(rtrain1$pred, rtrain1$real)
```

```
     RMSE  Rsquared       MAE
5.3609932 0.6996043 4.2877419
```

```r
rtest1 <- data.frame(real = test$Puntaje,
                     pred = predict(mod1, test))
postResample(rtest1$pred, rtest1$real)
```

```
     RMSE  Rsquared       MAE
5.6912636 0.7251652 4.5640581
```

Modelo 2

Modelo sin interacción entre estudio y curso, y efecto cúbico del sueño

```r
mod2 <- lm(Puntaje ~ Estudio + Curso + poly(Sueno,3), data = train)
mod2 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio + Curso + poly(Sueno, 3), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2223  -3.5103   0.2144   3.7144  15.0028

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.68071    0.77347  82.331  < 2e-16 ***
Estudio           1.65267    0.05063  32.642  < 2e-16 ***
CursoOnline      -2.30355    0.65931  -3.494 0.000521 ***
CursoPresencial  -0.84042    0.59308  -1.417 0.157124
poly(Sueno, 3)1  13.61664    5.42657   2.509 0.012432 *
poly(Sueno, 3)2 -13.97608    5.41673  -2.580 0.010176 *
poly(Sueno, 3)3   2.33371    5.41422   0.431 0.666641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.406 on 473 degrees of freedom
Multiple R-squared:  0.699, Adjusted R-squared:  0.6952
F-statistic: 183.1 on 6 and 473 DF,  p-value: < 2.2e-16
```

```r
mod2 |> car::vif()
```

```
                  GVIF Df GVIF^(1/(2*Df))
Estudio       1.003278  1        1.001638
Curso         1.013879  2        1.003452
poly(Sueno, 3) 1.014962  3        1.002478
```

```r
rtrain2 <- data.frame(real = train$Puntaje,
                      pred = predict(mod2, train))
postResample(rtrain2$pred, rtrain2$real)
```

```
     RMSE  Rsquared       MAE
5.3661831 0.6990224 4.2948581
```

```r
rtest2 <- data.frame(real = test$Puntaje,
                     pred = predict(mod2, test))
postResample(rtest2$pred, rtest2$real)
```

```
     RMSE  Rsquared       MAE
5.7066566 0.7233091 4.5616550
```

Modelo 3

Modelo con interacción entre estudio y curso, y efecto cuadrático del sueño

```r
mod3 <- lm(Puntaje ~ Estudio*Curso + poly(Sueno,2), data = train)
mod3 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio * Curso + poly(Sueno, 2), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9752  -3.4993   0.1593   3.7355  15.0056

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             64.46173    1.34178  48.042   <2e-16 ***
```

```
Estudio                   1.58900    0.10252  15.499   <2e-16 ***
CursoOnline              -3.86026    1.78846  -2.158   0.0314 *
CursoPresencial          -1.48059    1.66415  -0.890   0.3741
poly(Sueno, 2)1          13.83246    5.43308   2.546   0.0112 *
poly(Sueno, 2)2         -13.90495    5.42037  -2.565   0.0106 *
Estudio:CursoOnline       0.12970    0.13727   0.945   0.3452
Estudio:CursoPresencial   0.05168    0.12737   0.406   0.6851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.407 on 472 degrees of freedom
Multiple R-squared:  0.6995,    Adjusted R-squared:  0.695
F-statistic:   157 on 7 and 472 DF,  p-value: < 2.2e-16
```

```
mod3 |> car::vif()
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

```
                   GVIF Df GVIF^(1/(2*Df))
Estudio         4.111396  1        2.027658
Curso          52.587797  2        2.692906
poly(Sueno, 2)  1.014527  2        1.003612
Estudio:Curso  70.912155  2        2.901885
```

```
rtrain3 <- data.frame(real = train$Puntaje,
                      pred = predict(mod3, train))
postResample(rtrain3$pred, rtrain3$real)
```

```
     RMSE  Rsquared       MAE
5.3619345 0.6994988 4.2835598
```

```
rtest3 <- data.frame(real = test$Puntaje,
                     pred = predict(mod3, test))
postResample(rtest3$pred, rtest3$real)
```

```
     RMSE  Rsquared       MAE
5.6983610 0.7243944 4.5615552
```

Modelo 4

Modelo sin interacción entre estudio y curso, y efecto cuadrático del sueño

```
mod4 <- lm(Puntaje ~ Estudio + Curso + poly(Sueno,2), data = train)
mod4 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio + Curso + poly(Sueno, 2), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-15.235  -3.520   0.204   3.663  15.096

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.68627    0.77270  82.421  < 2e-16 ***
Estudio           1.65206    0.05057  32.671  < 2e-16 ***
CursoOnline      -2.29362    0.65835  -3.484  0.00054 ***
CursoPresencial  -0.84263    0.59254  -1.422  0.15567
poly(Sueno, 2)1  13.60589    5.42185   2.509  0.01242 *
poly(Sueno, 2)2 -13.97481    5.41207  -2.582  0.01012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.401 on 474 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6957
F-statistic:   220 on 5 and 474 DF,  p-value: < 2.2e-16
```

```
mod4 |> car::vif()
```

```
                   GVIF Df GVIF^(1/(2*Df))
Estudio        1.002495  1        1.001247
Curso          1.011553  2        1.002876
poly(Sueno, 2) 1.011788  2        1.002934
```

```
rtrain4 <- data.frame(real = train$Puntaje,
                      pred = predict(mod4, train))
postResample(rtrain4$pred, rtrain4$real)
```

```
     RMSE  Rsquared       MAE
5.3672369 0.6989041 4.2917029
```

```
rtest4 <- data.frame(real = test$Puntaje,
                     pred = predict(mod4, test))
postResample(rtest4$pred, rtest4$real)
```

```
     RMSE  Rsquared       MAE
5.7150810 0.7223909 4.5595529
```

Modelo 5

Modelo con interacción entre estudio y curso, y efecto lineal del sueño

```
mod5 <- lm(Puntaje ~ Estudio*Curso + Sueno, data = train)
mod5 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio * Curso + Sueno, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4766  -3.5850   0.0255   3.6317  15.1413

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             60.48361    2.01443  30.025   <2e-16 ***
Estudio                  1.59405    0.10310  15.461   <2e-16 ***
CursoOnline             -3.79867    1.79881  -2.112   0.0352 *
CursoPresencial         -1.34008    1.67303  -0.801   0.4235
Sueno                    0.54329    0.21508   2.526   0.0119 *
Estudio:CursoOnline      0.13243    0.13808   0.959   0.3380
Estudio:CursoPresencial  0.04604    0.12810   0.359   0.7194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.439 on 473 degrees of freedom
Multiple R-squared:  0.6953,    Adjusted R-squared:  0.6914
F-statistic: 179.9 on 6 and 473 DF,  p-value: < 2.2e-16
```

```r
mod5 |> car::vif()
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

```
                  GVIF Df GVIF^(1/(2*Df))
Estudio       4.109881  1        2.027284
Curso        52.527316  2        2.692132
Sueno         1.009598  1        1.004788
Estudio:Curso 70.854588  2        2.901296
```

```r
rtrain5 <- data.frame(real = train$Puntaje,
                      pred = predict(mod5, train))
postResample(rtrain5$pred, rtrain5$real)
```

```
    RMSE Rsquared      MAE
5.399184 0.695309 4.321880
```

```r
rtest5 <- data.frame(real = test$Puntaje,
                     pred = predict(mod5, test))
postResample(rtest5$pred, rtest5$real)
```

```
     RMSE  Rsquared       MAE
5.8253582 0.7095261 4.6183781
```

Modelo 6

Modelo sin interacción entre estudio y curso, y efecto lineal del sueño

```r
mod6 <- lm(Puntaje ~ Estudio + Curso + Sueno, data = train)
mod6 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio + Curso + Sueno, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-14.7264  -3.5649   0.0758   3.7500  15.1672

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     59.79454    1.71081  34.951  < 2e-16 ***
Estudio          1.65543    0.05085  32.555  < 2e-16 ***
CursoOnline     -2.19989    0.66125  -3.327 0.000946 ***
CursoPresencial -0.77005    0.59540  -1.293 0.196523
Sueno            0.53398    0.21465   2.488 0.013199 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.433 on 475 degrees of freedom
Multiple R-squared:  0.6947,    Adjusted R-squared:  0.6921
F-statistic: 270.2 on 4 and 475 DF,  p-value: < 2.2e-16
```

```
mod6 |> car::vif()
```

```
           GVIF Df GVIF^(1/(2*Df))
Estudio 1.001824  1        1.000912
Curso   1.008037  2        1.002003
Sueno   1.007690  1        1.003838
```

```
rtrain6 <- data.frame(real = train$Puntaje,
                      pred = predict(mod6, train))
postResample(rtrain6$pred, rtrain6$real)
```

```
     RMSE  Rsquared       MAE
5.4048542 0.6946688 4.3300562
```

```
rtest6 <- data.frame(real = test$Puntaje,
                     pred = predict(mod6, test))
postResample(rtest6$pred, rtest6$real)
```

```
     RMSE  Rsquared       MAE
5.8482996 0.7068482 4.6265608
```

Resumiendo:

		RMSE		MAE		R^2		Multicolinealidad	Significancia
Modelo		train	test	train	test	train	test		
1	Puntaje ~ Estudio*Curso + poly(Sueno,3)	5,361	5,691	4,288	4,564	0,7	0,725	Sin problemas	efecto cúbico del sueño no es significativo, la interacción estudio x curso no es significativa
			6,16%		6,44%		0,025		
2	Puntaje ~ Estudio + Curso + poly(Sueno,3)	5,366	5,707	4,295	4,562	0,699	0,723	Sin problemas	efecto cúbico del sueño no es significativo
			6,35%		6,22%		0,024		
3	Puntaje ~ Estudio*Curso + poly(Sueno,2)	5,362	5,698	4,284	4,562	0,699	0,724	Sin problemas	interacción estudio x curso no es significativa
			6,27%		6,49%		0,025		
4	Puntaje ~ Estudio + Curso + poly(Sueno,2)	5,367	5,715	4,292	4,56	0,699	0,722	Sin problemas	todas las variables aportan
			6,48%		6,24%		0,023		
5	Puntaje ~ Estudio*Curso + Sueno	5,399	5,826	4,322	4,618	0,695	0,71	Sin problemas	interacción estudio x curso no es significativa
			7,91%		6,85%		0,015		
6	Puntaje ~ Estudio + Curso + Sueno	5,405	5,848	4,33	4,627	0,695	0,707	Sin problemas	todas las variables aportan
			8,20%		6,86%		0,012		

Los modelos 4 y 6 son los ganadores, tienen mejores indicadores y son los más sencillos.

Comencemos a compararlos con criterios estadísticos:

```
mod4 |> residuals() |> shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  residuals(mod4)
W = 0.99753, p-value = 0.7041
```

```
mod6 |> residuals() |> shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  residuals(mod6)
W = 0.99733, p-value = 0.6367
```

```
mod4 |> car::ncvTest()
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.7347918, Df = 1, p = 0.39133
```

```
mod6 |> car::ncvTest()
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.7384256, Df = 1, p = 0.39017
```

```
mod4 |> lmtest::dwtest()
```

```
    Durbin-Watson test

data:  mod4
DW = 1.9472, p-value = 0.2792
alternative hypothesis: true autocorrelation is greater than 0
```

```
mod6 |> lmtest::dwtest()
```

```
    Durbin-Watson test

data:  mod6
DW = 1.94, p-value = 0.2534
alternative hypothesis: true autocorrelation is greater than 0
```

```
mod4 |> AIC()
```

```
[1] 2989.282
```

```
mod6 |> AIC()
```

```
[1] 2993.987
```

```
summary(mod4)$adj.r.squared
```

```
[1] 0.695728
```

```
summary(mod6)$adj.r.squared
```

```
[1] 0.6920976
```

Finalmente, el criterio de AIC nos sugiere utilizar el modelo 4:

Modelo sin interacción entre estudio y curso, y efecto cuadrático del sueño

```r
mod4 <- lm(Puntaje ~ Estudio + Curso + poly(Sueno,2), data = train)
mod4 |> summary()
```

```
Call:
lm(formula = Puntaje ~ Estudio + Curso + poly(Sueno, 2), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-15.235  -3.520   0.204   3.663  15.096

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.68627    0.77270  82.421  < 2e-16 ***
Estudio           1.65206    0.05057  32.671  < 2e-16 ***
CursoOnline      -2.29362    0.65835  -3.484  0.00054 ***
CursoPresencial  -0.84263    0.59254  -1.422  0.15567
poly(Sueno, 2)1  13.60589    5.42185   2.509  0.01242 *
poly(Sueno, 2)2 -13.97481    5.41207  -2.582  0.01012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.401 on 474 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6957
F-statistic:   220 on 5 and 474 DF,  p-value: < 2.2e-16
```