Análisis de regresión

Capítulo 1: Análisis de correlación y regresión lineal

Mg. Sc. J. Eduardo Gamboa U.

Presentación

Datos del docente

Mg. Jesús Eduardo Gamboa Unsihuay

Correo: jgamboa@lamolina.edu.pe

Datos del curso

EP6003 Análisis de regresión

Link de la sesión: Teams

Repositorio: GitHub

Sumilla

El curso Análisis de Regresión pertenece al área de formación de la especialidad, es de carácter obligatorio y de naturaleza teórico-práctica. El curso pretende desarrollar en el estudiante la capacidad de modelar una variable cuantitativa en función de otras y realizar inferencia sobre sus parámetros. Comprende las siguientes unidades o capítulos: Análisis de correlación y regresión lineal simple. Modelo de Regresión múltiple. Comprobación de la adecuación del modelo. Transformaciones para corregir inadecuaciones. Valores atípicos e influenciales. Variables indicadoras. Métodos de selección de variables. Multicolinealidad. Modelos de egresión polinomiales. Regresión y fundamentos del aprendizaje automático.

Evaluaciones

- Práctica Calificada 1:
  - Evalúa Unidad 1 y 2
  - Semana 2
  - Teórica - **práctica**
  - Ponderación: 20%
- Evaluación 1:
  - Evalúa Unidad 1, 2, 3 y 4
  - Semana 3
  - **Teórico** - práctico
  - Ponderación: 25%
- Práctica Calificada 2:
  - Evalúa Unidad 5, 6 y 7
  - Semana 4
  - Teórica - **práctica**
  - Ponderación: 20%

Evaluaciones

- Evaluación 2:
  - Evalúa Unidad 5, 6, 7, 8, 9 y 10
  - Semana 4
  - **Teórico** - práctico
  - Ponderación: 25%
- Actitudinal
  - Durante todas las semanas del curso
  - Participación, asistencia y entrega de tareas
  - Ponderación: 10%

Unidades del curso

1. Análisis de correlación y regresión lineal simple
2. Análisis de regresión lineal múltiple
3. Supuestos y comprobación de la adecuación del modelo
4. Transformaciones para corregir inadecuaciones del modelo
5. Valores atípicos e influenciales
6. Variables indicadoras
7. Selección de variables
8. Modelos polinomiales
9. Multicolinealidad
10. (Regresión y fundamentos del aprendizaje automático)

Correlación

- Entre dos o más variables pueden existir relaciones de asociación o de causa–efecto.
- La correlación analiza asociación, no implica causalidad.
- La asociación entre variables cuantitativas se estudia mediante gráficos de dispersión.
- Se asume que las variables en estudio son aleatorias o estocásticas.

- El diagrama de dispersión es una herramienta gráfica que permite visualizar la distribución conjunta de dos variables cuantitativas.
- Cuando se evalúa una posible relación causal, la variable independiente se representa usualmente en el eje horizontal (abscisas) y la variable respuesta en el eje vertical (ordenadas).
- En presencia de observaciones repetidas en un mismo punto, puede emplearse la técnica de *jittering* para mejorar la visualización.

```r
datos = read.csv('U1_datos_1.csv')
x = datos$horas_estudio
y = datos$puntaje_examen
```

```r
library(ggplot2)

ggplot(datos, aes(x = horas_estudio, y = puntaje_examen)) +
  geom_point(size = 2) +
  labs(title = "Diagrama de dispersión",
       subtitle = "Horas de estudio vs. Puntaje en examen",
       x = "Horas de estudio semanal",
       y = "Puntaje en examen") +
  theme_minimal()
```

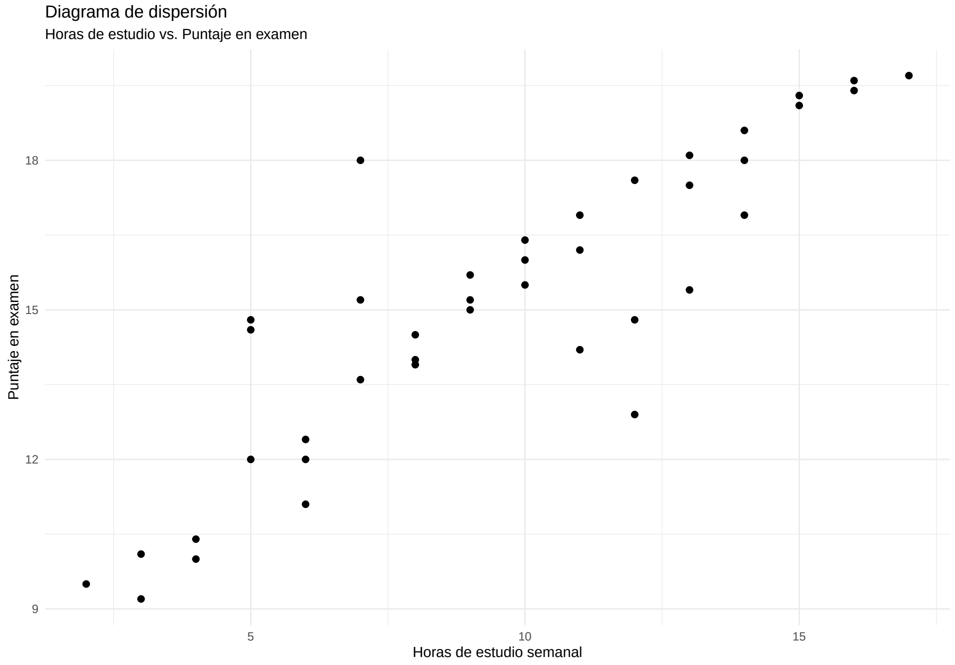


Diagrama de dispersión
Horas de estudio vs. Puntaje en examen
Puntaje en examen
18
15
12
9
5
10
15
Horas de estudio semanal

Coeficiente de correlación de Pearson

- Mide la asociación lineal entre dos variables cuantitativas.
- Su valor varía entre -1 y 1.
- Es más adecuado cuando las variables siguen (aproximadamente) una distribución normal.

Estimación puntual

El parámetro de correlación poblacional se denota por ρ, y su estimador muestral por r o $\hat{\rho}$.
Dado un par de variables X y Y, la estimación puntual de ρ está dada por:

$$r = \hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

```r
Num = sum((x-mean(x))*(y-mean(y)))
Den = sqrt(sum((x-mean(x))^2))*sqrt(sum((y-mean(y))^2))
(r = Num/Den)
```

```
[1] 0.8770203
```

```r
cor(x,y)
```

```
[1] 0.8770203
```

La asociación entre las horas de estudio y el puntaje en el examen es muy alta.

Estimación intervalar

La distribución muestral de r es asimétrica, por lo que no puede asumirse normalidad directamente.

Sin embargo, si X e Y siguen una distribución normal bivariada con coeficiente de correlación ρ, la transformación de Fisher:

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

sigue una distribución normal:

$$Z_r \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

Un intervalo de confianza $(1-\alpha)100\%$ para Z_ρ es:

$$LI(Z_\rho) = Z_r - Z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}, \quad LS(Z_\rho) = Z_r + Z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}$$

La transformación inversa permite obtener el intervalo para ρ:

$$\rho = \frac{e^{2Z}-1}{e^{2Z}+1}$$

```r
Zr = 0.5*log((1+r)/(1-r))
n  = nrow(datos)
li = Zr - qnorm(0.975)*sqrt(1/(n-3))
ls = Zr + qnorm(0.975)*sqrt(1/(n-3))
LI = (exp(2*li)-1)/(exp(2*li)+1)
LS = (exp(2*ls)-1) /(exp(2*ls)+1)
c(LI, LS)
```

```
[1] 0.7780833 0.9334980
```

```r
cor.test(x,y, method = "pearson")$conf.int
```

```
[1] 0.7780833 0.9334980
attr(,"conf.level")
[1] 0.95
```

Prueba de hipótesis

Se desea contrastar:

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0$$

El estadístico de prueba es:

$$t_{\text{calc}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

el cual sigue una distribución t de Student con $n-2$ grados de libertad.
El *p-valor* se calcula como:

$$p\text{-valor} = 2P(t_{n-2} > |t_{\text{calc}}|)$$

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0 \qquad \alpha = 0.05$$

```r
tcalc = r*sqrt(n-2)/sqrt(1-r**2)
(pvalor = 2*pt(abs(tcalc), df = n-2, lower.tail = FALSE))
```

```
[1] 1.163391e-13
```

```r
cor.test(x,y, method = "pearson")
```

```

	Pearson's product-moment correlation

data:  x and y
t = 11.253, df = 38, p-value = 1.163e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7780833 0.9334980
sample estimates:
      cor 
0.8770203 
```

Regresión

La regresión es una técnica estadística que modela la relación entre dos o más variables.

Según Montgomery et al. (2012), permite analizar la relación de dependencia o influencia de una o más variables explicativas sobre una variable respuesta.

Weisberg (2014) enfatiza su utilidad para explicar patrones sistemáticos presentes en los datos.

En el contexto más simple, se consideran dos variables:

- Y: variable respuesta, dependiente, objetivo o endógena.
- X: variable predictora, independiente, explicativa o exógena.

En un contexto univariado, el interés estadístico suele centrarse en el estudio de la media de una variable aleatoria Y, definida como $\mu = E(Y)$, así como en su variabilidad mediante intervalos de confianza. En particular, un intervalo de confianza para la media poblacional está dado por:

$$IC(\mu) = \bar{y} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

Sin embargo, en muchas aplicaciones prácticas, la variable Y puede verse influenciada por otra variable X, desplazando el interés hacia el estudio de la media condicional $E(Y \mid X)$. Esto plantea interrogantes adicionales sobre las características que deben cumplir X e Y y sobre cómo se modifica el proceso de inferencia estadística.

Para el análisis de regresión se asume que:

$$Y = f(X, \varepsilon)$$

donde ε representa un término de error aleatorio que recoge la variabilidad no explicada por el modelo.

Modelo de regresión lineal simple

Se define el modelo de regresión lineal simple como:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

donde:

- β_0 es el intercepto,
- β_1 es la pendiente,
- ε_i es el término de error aleatorio, con $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$.

Inferencia estadística: estimación puntual

Mínimos cuadrados ordinarios

El método de mínimos cuadrados busca estimar los parámetros del modelo minimizando la suma de los cuadrados de los residuos.

Máxima verosimilitud

Asumiendo que los errores siguen una distribución normal,

$$\varepsilon_i \sim N(0, \sigma^2)$$

se tiene que:

$$Y_i | \beta_0, \beta_1, \sigma^2 \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

La función de densidad de cada observación es:

$$f(Y_i | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right)^2 \right\}$$

La función de verosimilitud conjunta viene dada por:

$$L(\beta_0, \beta_1, \sigma^2 \mid Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f(Y_i | \beta_0, \beta_1, \sigma^2)$$

Los estimadores son:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde:

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- $\hat{\beta}_0$: valor estimado de Y cuando $X = 0$.
- $\hat{\beta}_1$: cambio esperado en Y ante un incremento unitario en X.

La interpretación de estos coeficientes debe realizarse con cautela, considerando el contexto del problema y el rango observado de la variable explicativa.

```r
sxx   = 1/n*sum((x-mean(x))**2)
sxy   = 1/n*sum((x-mean(x))*(y-mean(y)))
beta1 = sxy / sxx
beta0 = mean(y) - beta1*mean(x)
coef  = c(beta0,beta1)
names(coef) = c("beta0", "beta1")
coef
```

```
    beta0     beta1 
8.9900769 0.6413077 
```

```r
modelo = lm(puntaje_examen ~ horas_estudio, datos)
modelo |> coef()
```

```
  (Intercept) horas_estudio 
    8.9900769     0.6413077 
```

```r
modelo |> summary()
```

```

Call:
lm(formula = puntaje_examen ~ horas_estudio, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7858 -0.8200  0.1349  0.6463  4.5208 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.99008    0.58815   15.29  < 2e-16 ***
horas_estudio  0.64131    0.05699   11.25 1.16e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.453 on 38 degrees of freedom
Multiple R-squared:  0.7692,	Adjusted R-squared:  0.7631 
F-statistic: 126.6 on 1 and 38 DF,  p-value: 1.163e-13
```

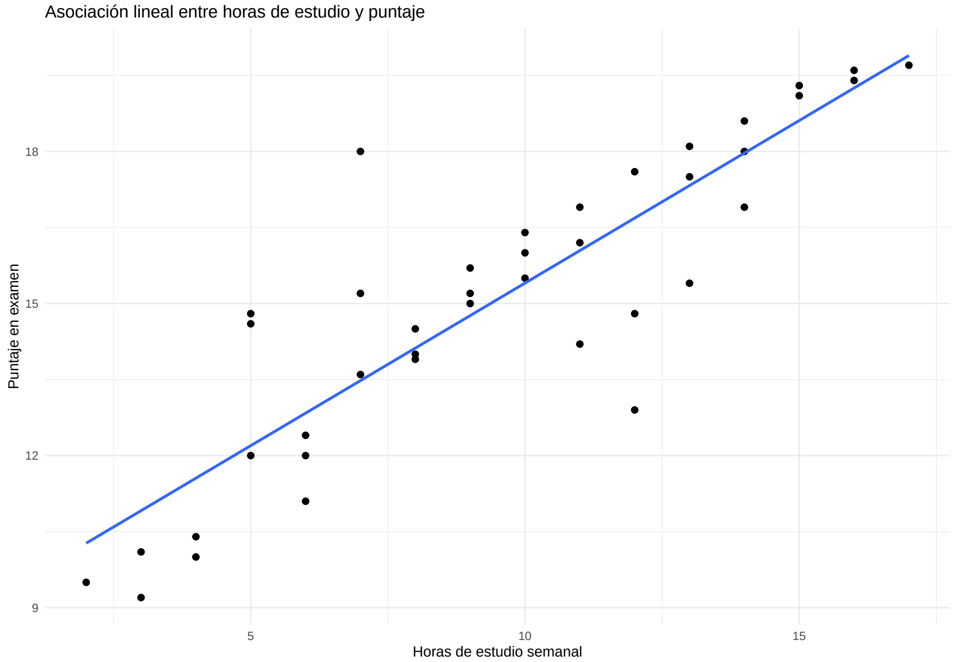
```r
library(broom)
modelo |> tidy()
```

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      8.99     0.588       15.3 8.13e-18
2 horas_estudio    0.641    0.0570      11.3 1.16e-13
```

$\hat{\beta}_0 = 8.99$ es el puntaje promedio en el examen cuando el alumno no estudia (cero horas de estudio)

$\hat{\beta}_1 = 0.64$: Por cada hora adicional de estudio, el puntaje promedio obtenido en el examen se incrementa en 0.64 puntos.

```r
ggplot(datos, aes(x = horas_estudio, y = puntaje_examen)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Asociación lineal entre horas de estudio y puntaje",
       x = "Horas de estudio semanal",
       y = "Puntaje en examen") +
  theme_minimal()
```


Asociación lineal entre horas de estudio y puntaje
Puntaje en examen
18
15
12
9
5
10
15
Horas de estudio semanal

Inferencia estadística: estimación intervalar

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ están sujetos a error muestral. Sus intervalos de confianza se construyen como:

$$IC(\beta_j) = \hat{\beta}_j \pm t_{1-\alpha/2,n-2}\, s_{\hat{\beta}_j}$$

donde $s_{\hat{\beta}_j}$ es el error estándar del estimador correspondiente.

```r
resumen = modelo |> tidy()
b0  = resumen$estimate[1]
b1  = resumen$estimate[2]
sb0 = resumen$std.error[1]
sb1 = resumen$std.error[2]
c(b0 - qt(0.99, n-2)*sb0, b0 + qt(0.99, n-2)*sb0)
```

```
[1]  7.561705 10.418449
```

```r
c(b1 - qt(0.99, n-2)*sb1, b1 + qt(0.99, n-2)*sb1)
```

```
[1] 0.5028980 0.7797174
```

```r
modelo |> confint(level = 0.98)
```

```
                     1 %       99 %
(Intercept)   7.561705 10.4184486
horas_estudio 0.502898  0.7797174
```

Con un 98% de confianza, se puede afirmar que el verdadero puntaje promedio cuando el alumno no estudia está contenido en el intervalo [7.56, 10.42] puntos.

Con un 98% de confianza, se puede afirmar que por cada hora adicional de estudio, el verdadero puntaje promedio se incrementa entre 0.50 y 0.78 puntos.

Inferencia: Pruebas de hipótesis

Para evaluar si la variable explicativa tiene influencia lineal sobre la variable respuesta, se plantea la hipótesis:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

Esta hipótesis puede contrastarse mediante el análisis de varianza (ANOVA). La variabilidad total se descompone en:

Fuente	GL	SC	CM
Regresión	1	SC_{Reg}	CM_{Reg}
Error	$n-2$	SC_{Error}	CM_{Error}
Total	$n-1$	SC_{Total}	

La hipótesis nula se rechaza si $F_{calc} > F_{1-\alpha,1,n-2}$ lo cual es equivalente a un *p-valor* menor que α.

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0 \qquad \alpha = 0.05$$

```r
anova = aov(y ~ x) |> tidy()
anova
```

```
# A tibble: 2 x 6
  term         df sumsq meansq statistic   p.value
  <chr>     <dbl> <dbl>  <dbl>     <dbl>     <dbl>
1 x             1 267.   267.       127.  1.16e-13
2 Residuals    38  80.2   2.11        NA  NA
```

$$F_{calc} = 127 \qquad F_{tab} = F_{0.95,1,38} = 4.098$$

$$p - valor = 1.16 \times 10^{-13} \qquad \alpha = 0.05$$

Se rechaza H_0, en conclusión…

Es posible evaluar afirmaciones puntuales sobre los parámetros. Por ejemplo:

$$H_0 : \beta_1 \leq b \quad \text{vs.} \quad H_1 : \beta_1 > b$$

El estadístico de prueba es:

$$t_{calc} = \frac{\hat{\beta}_1 - b}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

Estimación de la variabilidad

El valor de σ^2 es estimado mediante el Cuadrado Medio del Error

```
# Varianza = sigma^2
summary(modelo)$sigma**2
```

```
[1] 2.11128
```

```
anova$meansq[2]
```

```
[1] 2.11128
```

```
# Desviación estándar = sigma
summary(modelo)$sigma
```

```
[1] 1.453024
```

```
anova$meansq[2] |> sqrt()
```

```
[1] 1.453024
```

Además:

$$IC(\sigma^2) = \left(\frac{SCE}{\chi^2_{1-\alpha/2, n-2}}, \frac{SCE}{\chi^2_{\alpha/2, n-2}} \right)$$

```r
anova$sumsq[2]/qchisq(0.975, n-2)
```

```
[1] 1.410105
```

```r
anova$sumsq[2]/qchisq(0.025, n-2)
```

```
[1] 3.506729
```

Coeficiente de determinación

- El coeficiente de determinación se define como:

$$R^2 = \frac{SC_{\text{Reg}}}{SC_{\text{Total}}}$$

  e indica el porcentaje de variabilidad de la variable respuesta explicado por la variable predictora.

- El coeficiente de determinación ajustado es:

$$R^2_{\text{aj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

  donde p es el número de variables independientes.

```r
rsq    = anova$sumsq[1]/(anova$sumsq[1] + anova$sumsq[2])
adjrsq = 1 - (1-rsq)*(n-1)/(n-1-1)
summary(modelo)$r.squared
```

```
[1] 0.7691646
```

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

```
[1] 0.76309
```

```r
glance(modelo)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.769         0.763  1.45      127. 1.16e-13     1  -70.7  147.  152.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Estimación y predicción

Estimación

La media estimada de Y para un valor dado de $X = x$ es:

$$\hat{\mu} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Su varianza estimada es:

$$\widehat{Var}(\hat{\mu}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SC_X} \right)$$

Estimación del puntaje promedio cuando la cantidad de horas de estudio es 8.

```r
x0 = 8
y0 = b0 + b1*x0
y0
```

```
[1] 14.12054
```

```r
modelo |> predict(newdata = data.frame(1, horas_estudio = x0))
```

```
       1
14.12054
```

```r
scx   = sum((x-mean(x))**2)
varmu = anova$meansq[2]*(1/n + (x0-mean(x))**2/scx)
li = y0 - qt(0.975, 38)*sqrt(varmu)
ls = y0 + qt(0.975, 38)*sqrt(varmu)
c(li,ls)
```

```
[1] 13.62429 14.61678
```

```r
modelo |> predict(newdata = data.frame(1, horas_estudio = x0),interval = "
```

```
       fit      lwr      upr
1 14.12054 13.62429 14.61678
```

$$IC(\mu|x = 8) = (13.62, 14.62)$$

Con un 95% de confianza, el verdadero puntaje promedio cuando la cantidad de horas de estudio es 8 está contenido en el intervalo (13.62, 14.62) puntos.

Predicción

La predicción para una nueva observación Y_0 viene dada por:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x$$

con varianza estimada:

$$\widehat{Var}(\hat{y}_0) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SC_X}\right)$$

Predicción del puntaje cuando la cantidad de horas de estudio es 8.

```r
x0 = 8
y0 = b0 + b1*x0
y0
```

```
[1] 14.12054
```

```r
modelo |> predict(newdata = data.frame(1, horas_estudio = x0))
```

```
       1
14.12054
```

```
scx   = sum((x-mean(x))**2)
varmu = anova$meansq[2]*(1 + 1/n + (x0-mean(x))**2/scx)
li = y0 - qt(0.975, 38)*sqrt(varmu)
ls = y0 + qt(0.975, 38)*sqrt(varmu)
c(li,ls)
```

```
[1] 11.13748 17.10360
```

```
modelo |> predict(newdata = data.frame(1, horas_estudio = x0),interval = "
```

```
       fit      lwr     upr
1 14.12054 11.13748 17.1036
```

$$IP(Y|x=8) = (11.14, 17.10)$$

Con un 95% de confianza, se predice que el puntaje obtenido por un alumno que estudió 8 horas está contenido en el intervalo (11.14, 17.10) puntos.