Clase 12 Regresión lineal

OCE 386 - Introducción al análisis de datos con R.

Dr. José A. Gallardo | Pontificia Universidad Católica de Valparaíso

26 October 2021

PLAN DE LA CLASE

1.- Introducción

- Regresión lineal ¿Qué es y para qué sirve?
- Correlación v/s causalidad.
- Repaso ecuación de regresión lineal.
- Repaso betas y causalidad.
- Interpretación Regresión lineal con R.

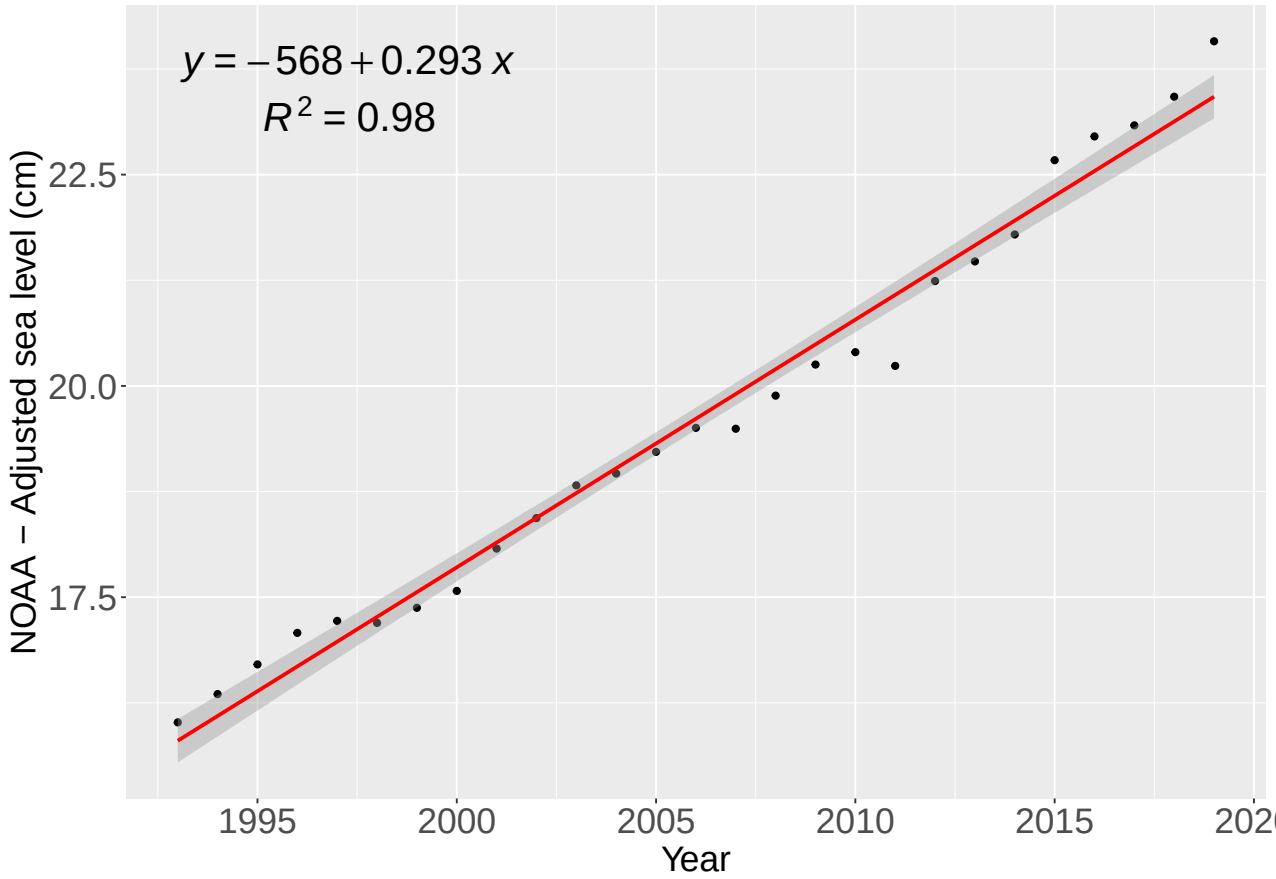
2.- Práctica con R y Rstudio cloud

- Realizar análisis de regresión lineal.
- Realizar gráficas avanzadas con ggplot2.
- Elaborar un reporte dinámico en formato pdf.

REGRESIÓN LINEAL

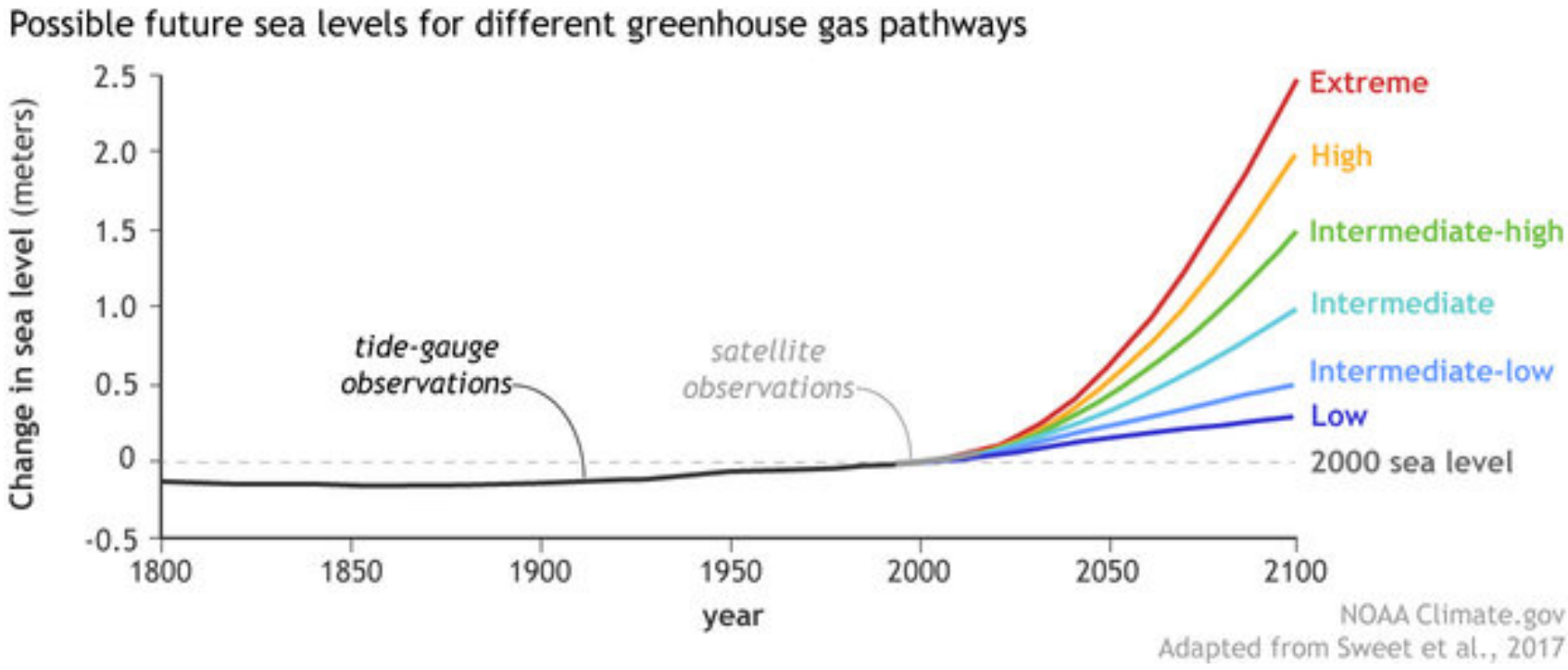
Herramienta estadística que permite determinar si existe una relación (asociación) entre una variable predictora (independiente) y la variable respuesta (dependiente).

Nivel del mar en función del tiempo. Fuente



REGRESIÓN LINEAL: PREDICCIÓN.

La ecuación de la regresión permite, bajo ciertos supuestos, predecir el valor de una variable respuesta "y" a partir de una o más variable predictoras "x".



REGRESIÓN LINEAL: BETAS

Betas miden la influencia del intercepto y la pendiente sobre la variable Y.

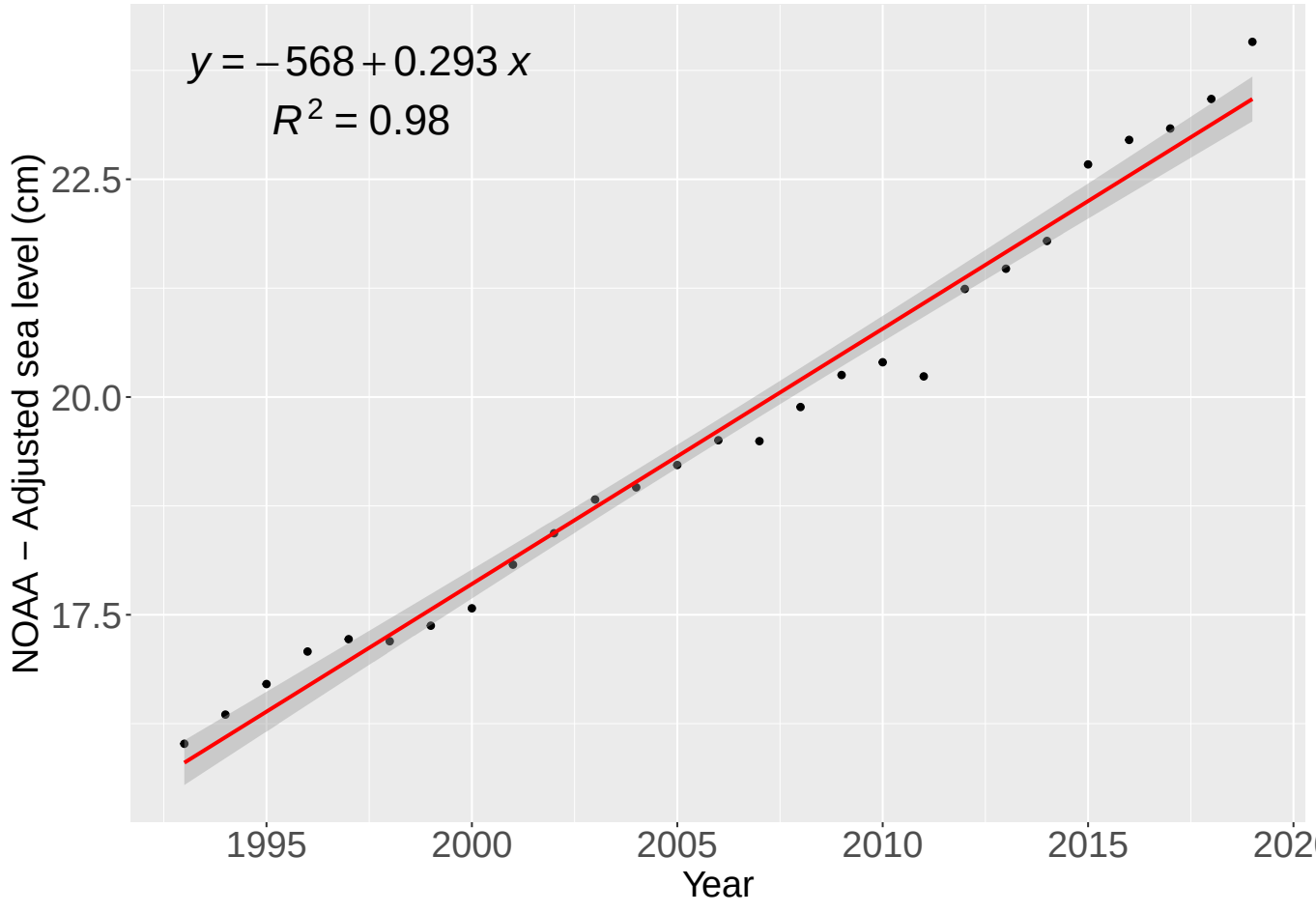
$Y = \beta_0 + \beta_1 X_1 + \epsilon$

β_0 = Intercepto = valor que toma “y” cuando x = 0.

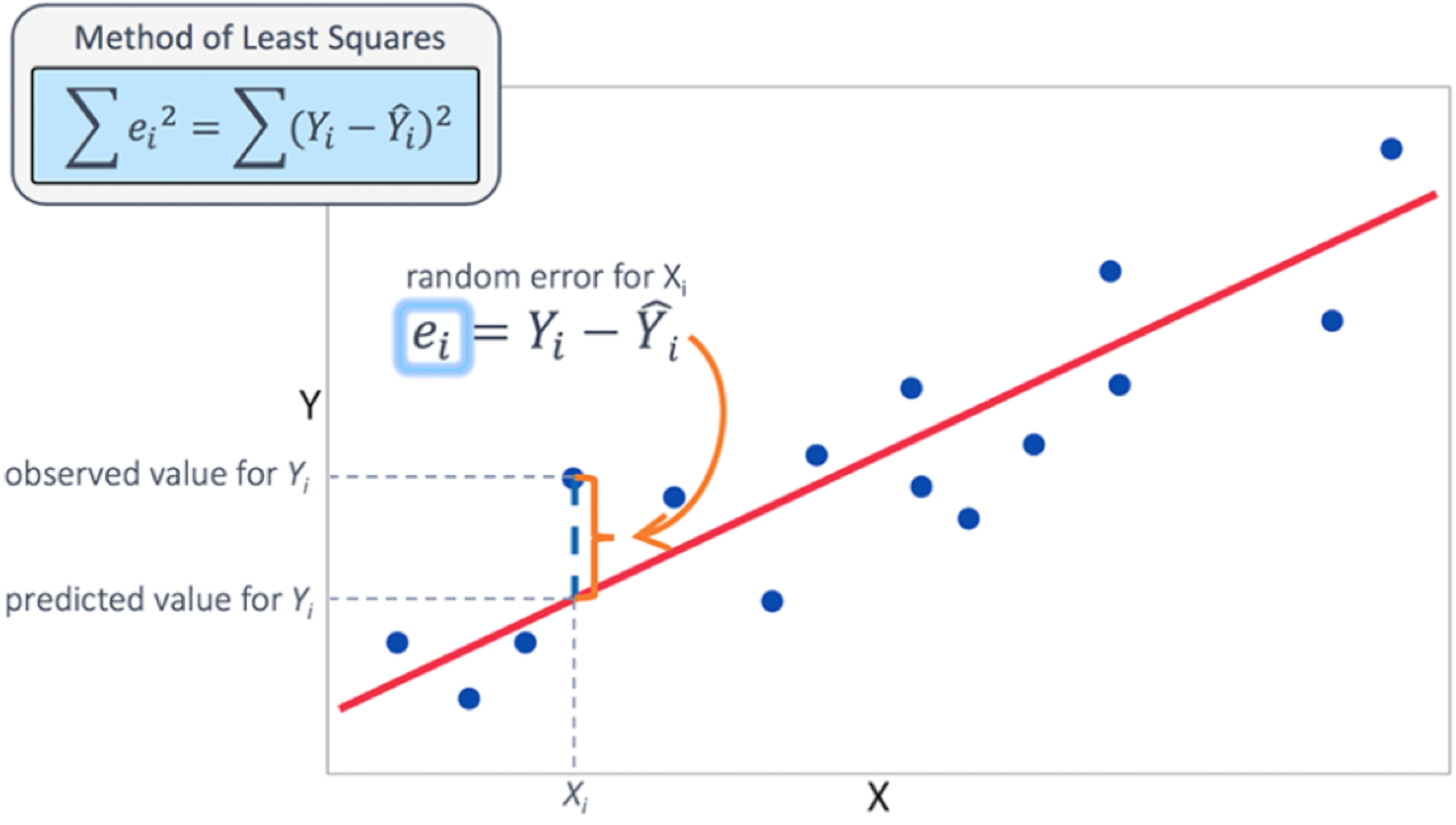
β_1 = Pendiente = Cambio promedio de “y” cuando “x” cambia en una unidad.

LINEA DE REGRESIÓN.

Línea de regresión: Corresponde a los valores "ajustados" o estimados de "y" en función de "x". Se calcula con los estimadores de *mínimos cuadrados* de $\beta_0 y \beta_1$



RESIDUOS Y MÉTODOS DE MÍNIMOS CUADRADOS



COCIENTE DE DETERMINACIÓN

R^2 mide la proporción de la variación muestral de "y" que es explicada por x (varía entre 0-1). Se calcula como el cuadrado del coeficiente de correlación de pearson.

R^2_{ajust} nos dice qué porcentaje de la variación de la variable dependiente es explicado por la o las variables independientes de manera conjunta.

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

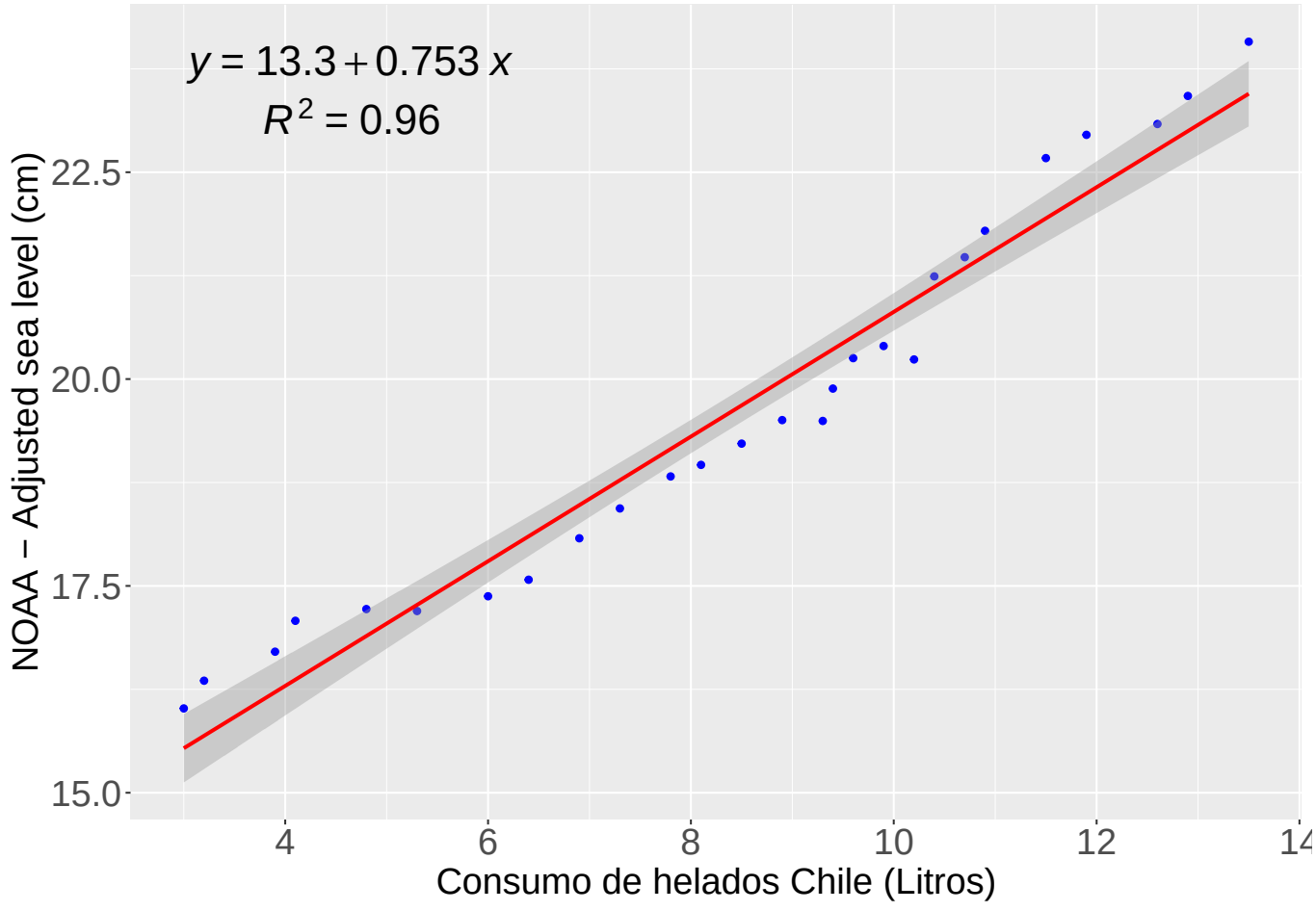
donde:

n = tamaño de la muestra

p = cantidad de variables predictoras en el modelo

CORRELACIÓN NO IMPLICA CAUSALIDAD

¿Si dejamos de tomar helados disminuirá el nivel del mar?
¿Qué factor "z" puede explicar la correlación entre consumo de helados y nivel del mar?



PRUEBAS DE HIPÓTESIS

Prueba de hipótesis del coeficiente de regresión y el intercepto **Tipo de prueba**: Prueba de t – student

La hipótesis nula en ambos casos es que los coeficiente (β_0) y (β_1) son iguales a 0, es decir sin asociación entre las variables.

$H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$

Prueba de hipótesis del modelo completo **Tipo de prueba**: Prueba de F.

La hipótesis nula es que los coeficientes son iguales a 0.

$H_0 : \beta_j = 0 \; ; \; j = 1, 2, ..., k$

REGRESIÓN LINEAL CON R: COEFICIENTES

```
reg <- lm(`NOAA - Adjusted sea level (cm)` ~ Year,
          data = sea_level_dat)
# summary(reg)
```

Coeficientes

	Estimate	Std. Error	t value	Pr(>)
(Intercept)	-568	1.650e+01	-34.45	<2e-16 ***
Year	0.29	8.223e-03	35.64	<2e-16 ***

REGRESIÓN LINEAL CON R: PRUEBA DE F

Anova de la regresión.

```
anova(reg) %>% kable()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	140.680295	140.6802949	1270.213	0
Residuals	25	2.768834	0.1107534	NA	NA

EXTRAER INFORMACIÓN DE LA REGRESIÓN LINEAL

```
summary(reg$residuals)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.84027 -0.18800 -0.07101   0.00000   0.25306   0.65581
```

```
summary(reg)$sigma
```

```
## [1] 0.3327963
```

```
summary(reg)$r.squared
```

```
## [1] 0.9806981
```

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

```
## [1] 0.9799261
```

PREDICCIÓN LINEAL DEL NIVEL DEL MAR

Predicción del nivel del mar promedio - próximos 3 años.

```
predict.lm(reg, newdata=data.frame(Year=c(2022,2023,2024)),
           interval="confidence")
```

```
##        fit      lwr      upr
## 1 24.30088 23.99951 24.60224
## 2 24.59394 24.27726 24.91062
## 3 24.88700 24.55485 25.21915
```

PREDICCIÓN LINEAL FUERA DEL RANGO OBSERVADO

Predicción del nivel del mar - en 3 mil años más.

¿Por qué esta predicción es inadecuada?

```
predict.lm(reg, newdata=data.frame(Year=c(5022,5023,5024)),
           interval="prediction")
```

```
##         fit      lwr      upr
## 1 903.4872 852.4058 954.5687
## 2 903.7803 852.6819 954.8787
## 3 904.0734 852.9580 955.1887
```

PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible en drive y Rstudio.cloud. **Clase_12**
- El trabajo práctico se realiza en Rstudio.cloud. **Guía 12 Regresión lineal**

RESUMEN DE LA CLASE

- **Elaborar hipótesis para una regresión lineal**
- **Realizar análisis de regresión lineal simple**
- **Interpretar coeficientes y realizar predicciones**