# Formas Funcionales (parte 1)

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyverse)
library(tidymodels)
```

## 1. Modelo Lineal Tradicional

Cuando la FRP de dos variables es de la forma:

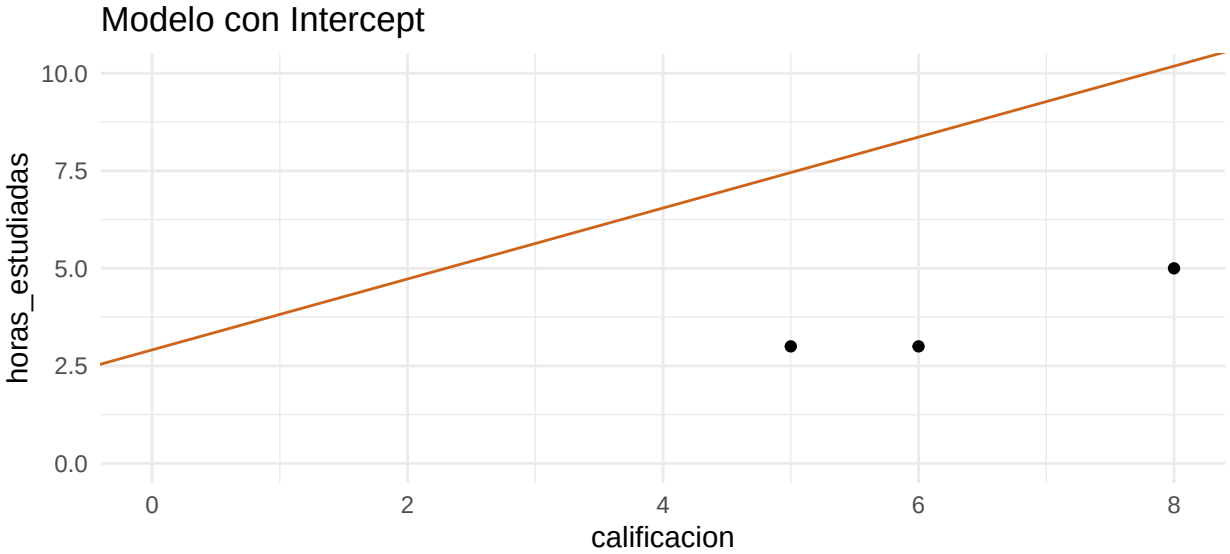
$$Y_i = \beta_0 + \beta_1 X_1 + e_i$$

La $\beta_1$ mide el cambio en $Y$ con respecto al cambio de $X$, es decir $\frac{\Delta Y}{\Delta X}$ (la cual conocemos como pendiente). Es importante notar que esto es dependiente de las escalas originales de las variables:

```
datos <- data.frame(horas_estudiadas = c(3,3,5,7), calificacion = c(5,6,8,9))

con_intercept <-
  linear_reg() %>%
  fit(calificacion~horas_estudiadas, data = datos)

datos %>% ggplot(aes(x = calificacion, y = horas_estudiadas))+
  geom_point()+
  geom_abline(intercept = con_intercept$fit$coefficients["(Intercept)"],
              slope = con_intercept$fit$coefficients["horas_estudiadas"], color = "chocolate3")+
  theme_minimal()+
  labs(title = "Modelo con Intercept")+
  ylim(0,10)+xlim(0, 8)
```




En este ejemplo el intercept es 2.9091 y la $\beta_1$ es 0.9091. Esto quiere decir que en promedio, un aumento de una hora estudiada refleja un aumento de 0.9091 en la calificación. Por otra parte, si se estudian 0 horas, la calificación esperada es de 2.9091.

## 2. Modelo Regresión a través del origen

Cuando la FRP de dos variables es de la forma:

$$Y_i = \beta_1 X_1 + e_i$$

Es decir el término del intercepto es ausente o es cero, se conoce como regresión a través del origen. La estimación de $\beta_1$ cambia a la siguiente fórmula:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

con

$$var\hat{\beta}_1 = \frac{\sigma^2}{\sum X_i^2}$$

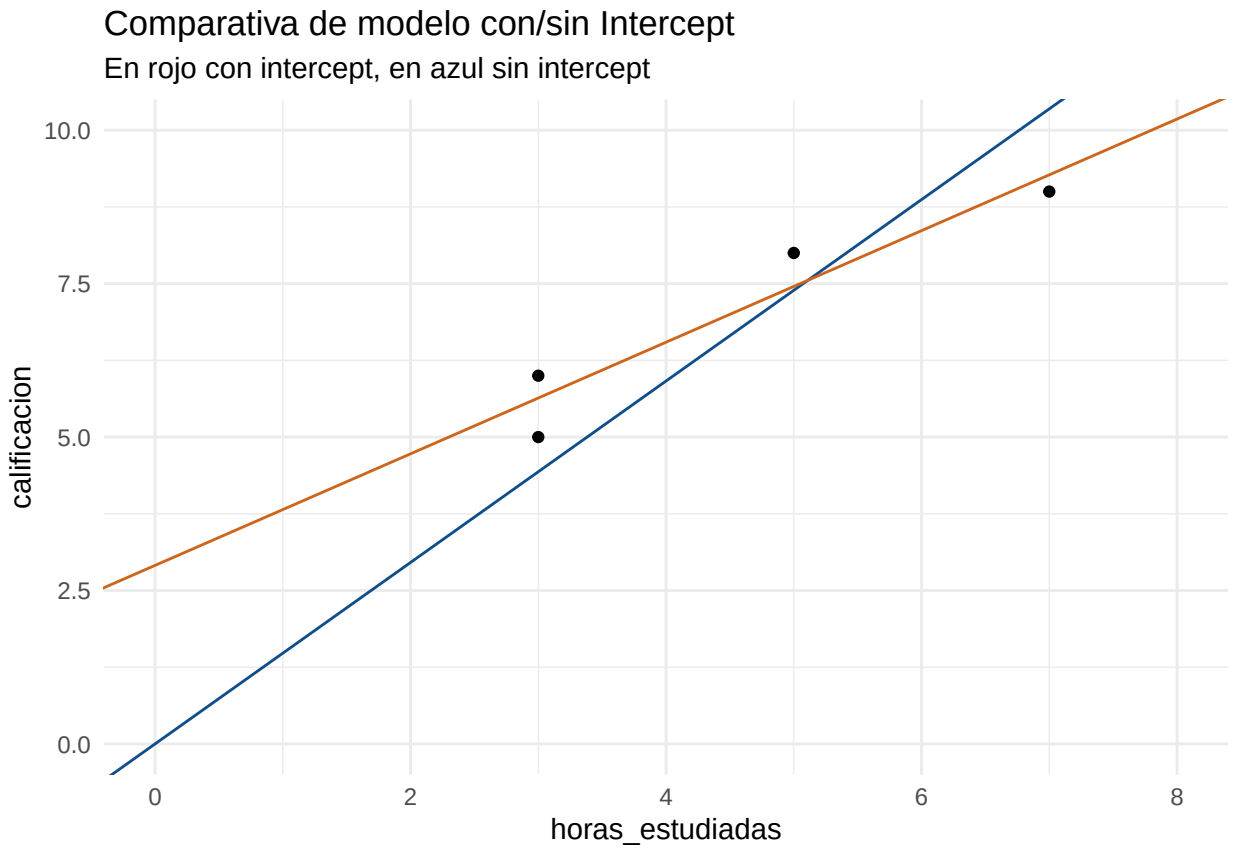
Al igual que en ejemplo de regresión lineal con intercepto, la varianza poblacional la estimamos con:

$$\hat{\sigma^2} = \frac{\sum \hat{e_i}^2}{n-1}$$

En este caso, hay que notar que el denominador esta $n-1$ en lugar de $n-2$.

```
sin_intercept <-
  linear_reg() %>%
  fit(calificacion~-1+horas_estudiadas, data = datos)

datos %>%
  ggplot(aes(x = horas_estudiadas, y = calificacion))+
  geom_point()+
  geom_abline(intercept = 0, slope = sin_intercept$fit$coefficients, color = "dodgerblue4")+
  geom_abline(intercept = con_intercept$fit$coefficients["(Intercept)"],
              slope = con_intercept$fit$coefficients["horas_estudiadas"], color = "chocolate3")+
  theme_minimal()+
  labs(title = "Comparativa de modelo con/sin Intercept",
       subtitle = "En rojo con intercept, en azul sin intercept")+
  ylim(0,10)+xlim(0, 8)
```




¿Qué podemos observar del ejemplo sin intercepto? En el modelo con intercepto, se observa que $\sum \hat{e}_i = 0$, pero en el caso de modelo sin intercepto no se cumple esto. Adicional, el coeficiente $R^2$ que es siempre no negativo en el modelo tradicional, puede volverse negativo en el modelo sin intercepto. Por esto, la $R^2$ convencional es poco recomendada de utilizar.

**$R^2$ a través del modelo de regresión en el origen** En el modelo sin intercepto, utilizamos lo que se conoce como $R^2$ **simple** el cual se define como:

$$R^2 simple = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2}$$

El $R^2 simple$ satisface que $0 < R^2 simple < 1$, pero no es comparable con la $R^2$ tradicional. Por este motivo comúnmente es recomendable utilizar el modelo con intercepto. Considere el caso donde se incluye el intercepto pero no es significativo.

## 3. Modelo sobre variables estandarizadas

En el modelo tradicional, las escalas de las variables influyen en la interpretación de los coeficientes de regresión. Esto se puede evitar si ambas variables (regresora y regresada) se expresan como variables estandarizadas (restar la media y dividir entre desviación estándar). De esta manera podemos reescribir a $X$ y $Y$ como:

$$X^* = \frac{X - \mu_x}{\sigma_x}$$
$$Y^* = \frac{Y - \mu_y}{\sigma_y}$$

(Al estar trabajando podemos ocupar los datos muestrales para estandarizar la variable). De esta manera podemos reescribir la regresión como:

$$Y_i^* = \beta_0^* + \beta_1^* X_1^* + e_i$$

Cuando se ajusta la regresión con variables estandarizadas, entonces el término de intercepto siempre es cero. Adicional, los coeficientes de regresión de las variables estandarizadas se conocen como **coeficientes beta**. La interpretación de los coeficientes beta es que si la regresora se incrementa en una desviación estándar, en promedio, la regresada aumenta en $\beta_1^*$ desviaciones estándar.

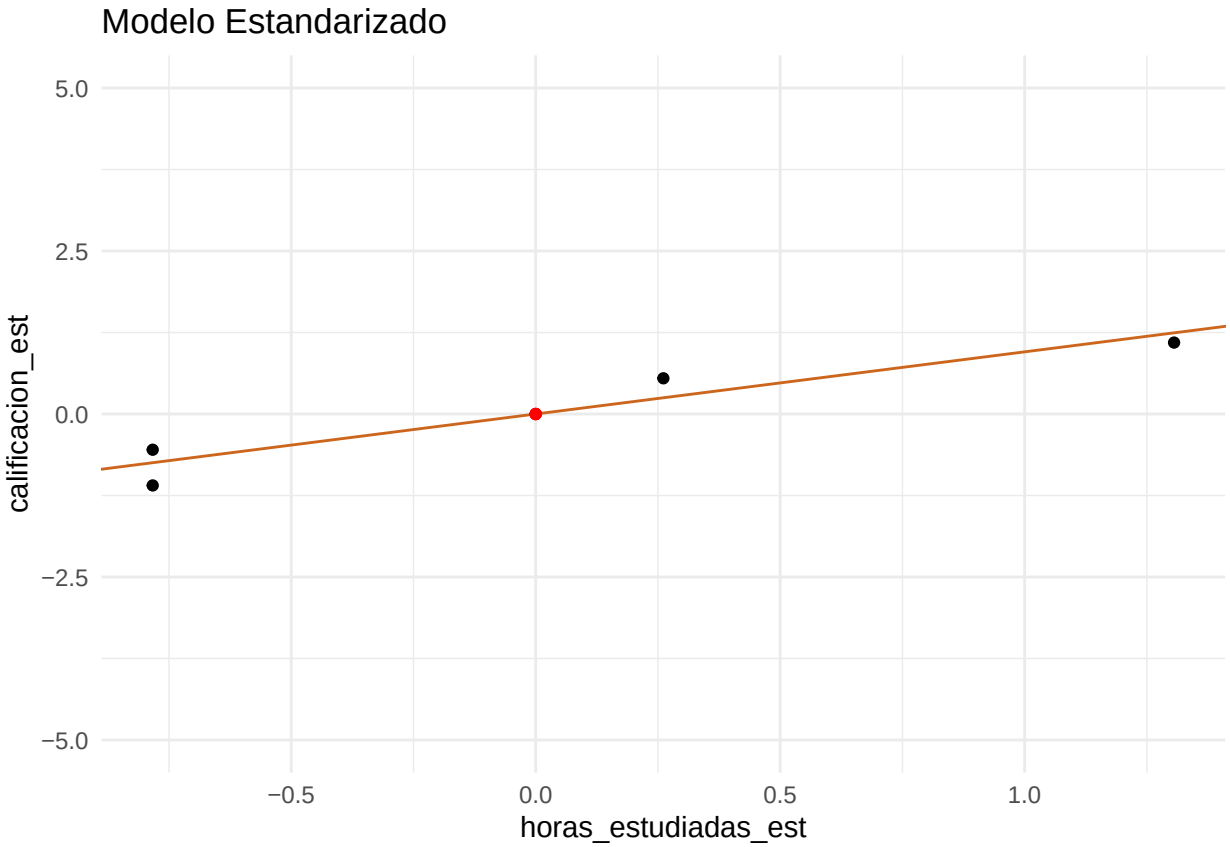
¿Cuál es la ventaja del modelo estandarizado sobre el modelo tradicional? Se puede ver mejor en el caso múltiple, pero en general, al tener varios coeficientes, podemos observar cuales tienen mayor impacto en una misma escala.

calificacion~horas_estudiadas

```
# modelo estandarizado
datos_est <-
  datos %>%
  mutate(horas_estudiadas_est = scale(horas_estudiadas, center = T, scale = T),
         calificacion_est = scale(calificacion, center = T, scale = T))

estandarizados <-
  linear_reg() %>%
  fit(calificacion_est~horas_estudiadas_est, data = datos_est)

datos_est %>%
  ggplot(aes(x = horas_estudiadas_est, y = calificacion_est))+
  geom_point()+
  geom_abline(intercept = estandarizados$fit$coefficients["(Intercept)"],
              slope = estandarizados$fit$coefficients["horas_estudiadas_est"], color = "chocolate3")+
  theme_minimal()+
  labs(title = "Modelo Estandarizado")+
  ylim(-5,5)+
  geom_point(x =0 , y = 0, color = "red")
```




Si se estandarizan las dos variables (es decir se centran y dividen entre su desviación estándar), entonces la recta pasa por el origen (ya no hay intercepto) y la interpretación cambia. En este caso lo interpretamos como un cambio en una desviación estándar de la variable independiente impacta en $\beta_1$ desviaciones estándar de la variable dependiente.

## 4. Ejercicio

Hay veces en las que ajustar una regresión lineal con datos crudos no es adecuado, por lo que se aplican transformaciones lineales y logaritmicas para poder interpretar el modelo. **Las transformaciones lineales no afectan el ajuste de un modelo de regresión y no afectan las predicciones**. Por otra parte, cambios en los inputs y coeficientes, **pueden mejorar la interpretabilidad** de los coeficientes y hacer el modelo más facil de interpretar.

Los coeficientes de regresión $\beta_j$ representan la **diferencia promedio** de $y$ cuando el predictor $x_i$ cambia en una unidad. Es por esto, que al hablar de escalas originales, nos podemos dar cuenta que el coeficiente está relacionado con la escala del regresor. Analicemos el siguiente ejercicio:

**Ejercicio 4.1: Se utilizarán datos de una encuensta de salarios en Estados Unidos que predice el salario basado en la altura de la persona (en pulgadas), para esto primero cargaremos los datos de la url:**

"https://raw.githubusercontent.com/Clark-Rhodes/INFO523/99c046debb9230fbfedaf08a67577a9a0c37e978/Intro-master/data/wages.csv"

**Ejercicio 4.2: Convierta la variable de height para tenerla en centímetros y en metros**

**Ejercicio 4.3: Realice la regresión lineal de earn~height_cm y earn~height_m, Interprete. ¿Qué observa de los coeficientes?, ¿Por qué sucede esto? ¿Hace sentido tener un intercept cuando height_cm = 0 o height_m = 0?. ¿Cuál es la $R^2$ del modelo?**

**Ejercicio 4.4: Centre las variables height_cm, height_m, ajuste nuevamente y observe los resultados. ¿El valor de $\beta_0$ cambió? ¿Cómo interpreta este nuevo valor? ¿El valor de $\beta_1$ cambió? ¿Cómo lo interpreta? Concluya**

**Ejercicio 4.5: Centre y escale las variables height_cm, height_m, ajuste nuevamente y observe los resultados. ¿El valor de $\beta_0$ cambió? ¿Cómo interpreta este nuevo valor? ¿El valor de $\beta_1$ cambió? ¿Cómo lo interpreta? Concluya**

**Ejercicio 4.6: Realize una tabla comparativa de los modelos y su interpretación de los coeficientes.**