

Formas Funcionales

1. Modelo Lineal Tradicional

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

Cuando la FRP de dos variables es de la forma:

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

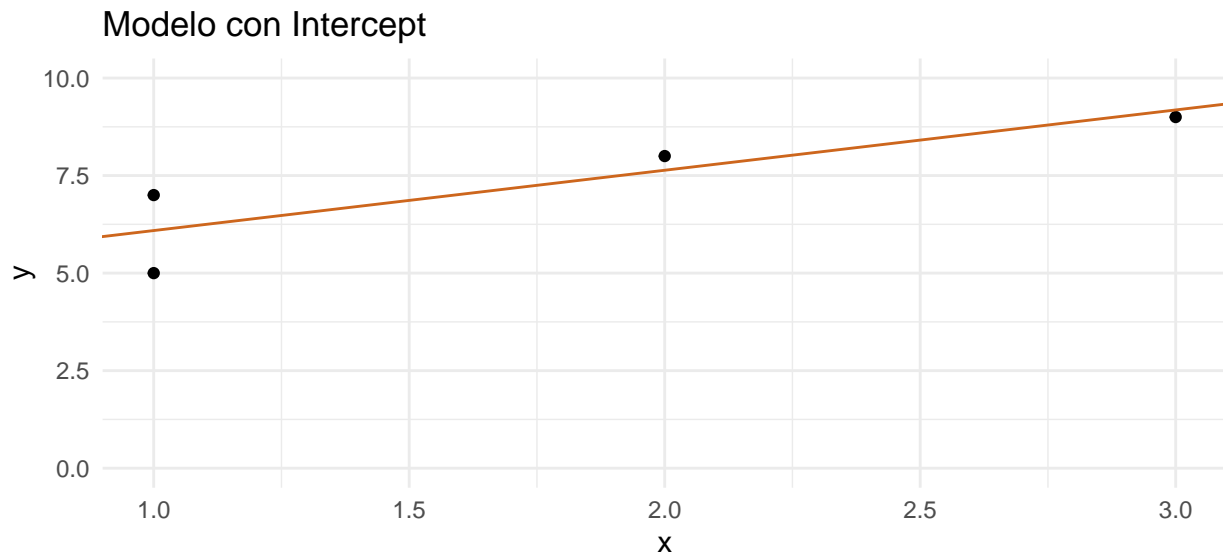
La β_1 mide el cambio en Y con respecto al cambio de X , es decir $\frac{\Delta Y}{\Delta X}$ (la cual conocemos como pendiente). Es importante notar que esto es dependiente de las escalas originales de las variables:

```
datos <- data.frame(x = c(1,1,2,3), y = c(5, 7,8,9))

con_intercept <-
  linear_reg() %>%
  fit(y~x, data = datos)

# beta1_con <- cov(datos$x, datos$y)/var(datos$x)
# beta0_con <- mean(datos$y)-beta1_con*mean(datos$x)

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



2. Modelo Regresión a través del origen

Cuando la FRP de dos variables es de la forma:

$$Y_i = \beta_1 X_i + e_i$$

Es decir el término del intercepto es ausente o es cero, se conoce como regresión a través del origen. La estimación de β_1 cambia a la siguiente fórmula:

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

con

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

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

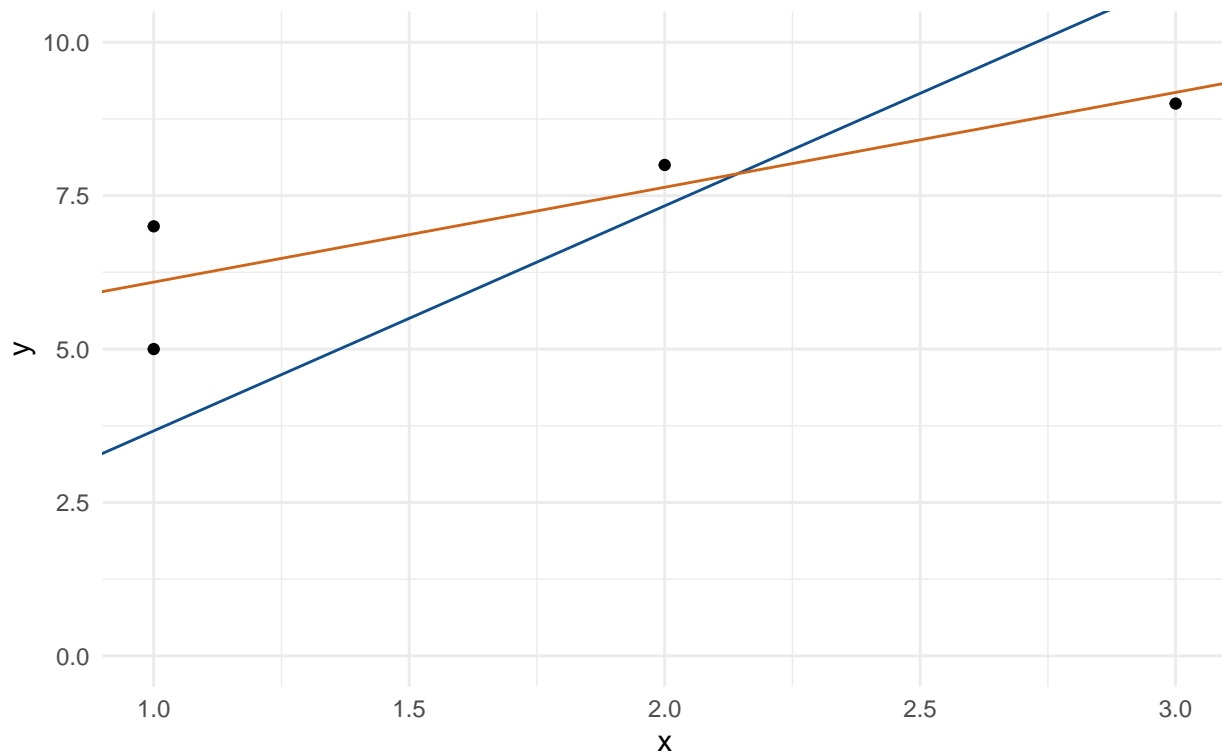
$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n - 1}$$

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

```
sin_intercept <-  
  linear_reg() %>%  
  fit(y~1+x, data = datos)  
  
beta1_sin <- sum(datos$x*datos$y)/sum(datos$x*datos$x)  
beta0_sin <- 0  
  
datos %>%  
  ggplot(aes(x = x, y = y))+  
  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["x"], color = "chocolate3")+  
  theme_minimal()+  
  labs(title = "Comparativa de modelo con/sin Intercept",  
        subtitle = "En rojo con intercept, en azul sin intercept")+  
  ylim(0,10)
```

Comparativa de modelo con/sin Intercept

En rojo con intercept, en azul sin intercept



¿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:

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

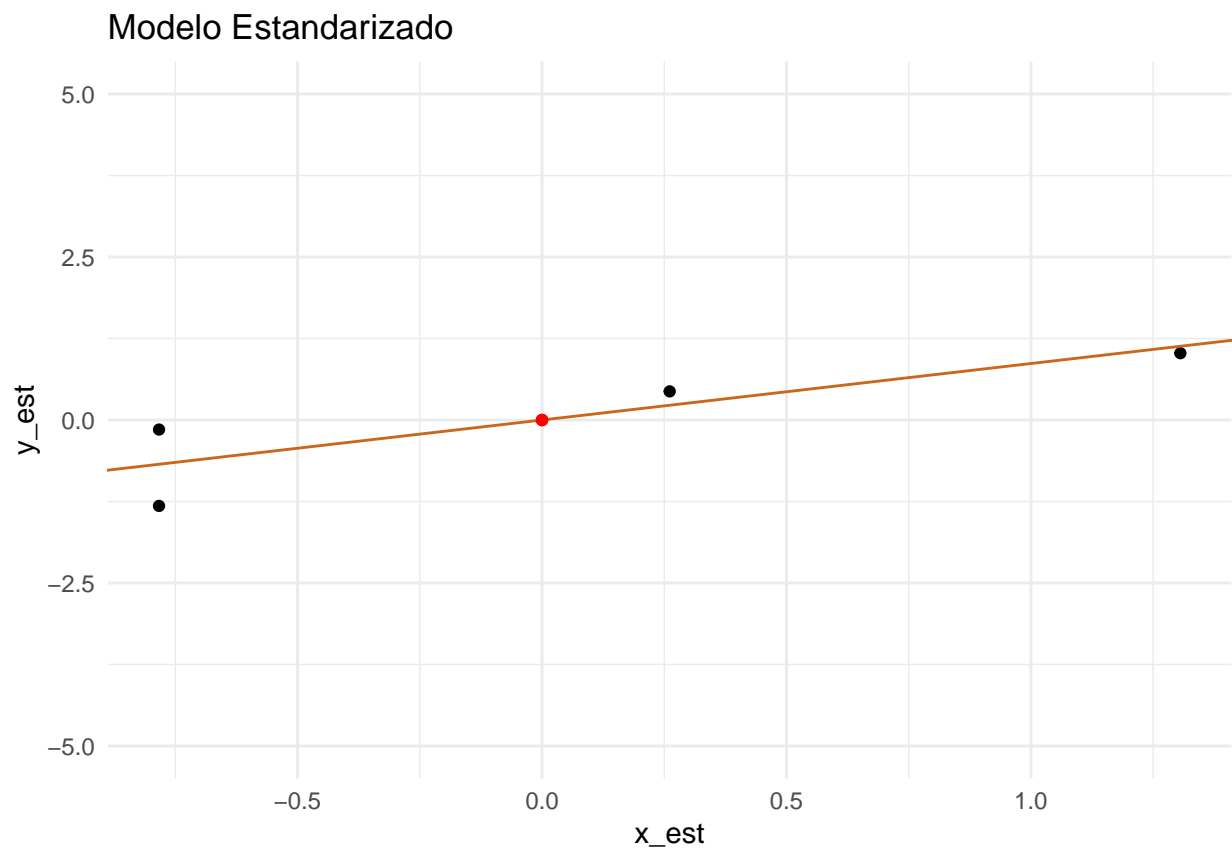
Cuando se ajusta la regresión con variables estandarizadas, entonces el término de intercepto siempre es cero. Adicional, los coeficientes de regresión de las variables estandarizadas se conocen como **coeficientes beta**. La interpretación de los coeficientes beta es que si la regresora se incrementa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

```
# modelo estandarizado
datos_est <-
  datos %>%
  mutate(x_est = scale(x, center = T, scale = T),
         y_est = scale(y, center = T, scale = T))

estandarizados <-
  linear_reg() %>%
  fit(y_est~x_est, data = datos_est)

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



4. Modelo Log-Log (Elasticidad constante)

Sea el modelo:

$$Y_i = \beta_0 X_1^{\beta_1} \exp^{e_i}$$

,

al aplicar logaritmo de ambos lados podemos transformarlo como:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 \ln(X_1) + e_i$$

En particular podemos reescribir a $\ln(\beta_0)$ como *alpha*, a $\ln(Y_i)$ como Y' y a $\ln(X_i)$ como X' y tendríamos un caso particular del modelo de regresión:

$$Y' = \alpha + \beta_1 X'_1 + e_i$$

Este modelo también recibe el nombre de modelo log-log, doble-log o log-lineal. Una característica atractiva del modelo log-log es que el coeficiente de la pendiente β_1 mide la elasticidad de Y con respecto de X, es decir: $\frac{\Delta\%Y}{\Delta\%X}$. Una característica especiales del modelo log-log es que el modelo supone que el coeficiente de la elasticidad entre Y y X se mantiene contante en el tiempo (elasticidad constante).

5. Modelo Log-Lin (Crecimiento exponencial)

Este es un modelo que comúnmente se llama semilogaritmico y busca medir la tasa de crecimiento. Suponga que el modelo sigue la forma:

$$Y_t = Y_0(1 + r)^t$$

donde r es la tasa de crecimiento compuesta de Y , es decir a través del tiempo. Al aplicar logaritmo natural:

$$\log(Y_t) = \log(Y_0) + t\log(1 + r)$$

Con $\beta_0 = \log(Y_0)$ y $\beta_1 = \log(1 + r)$, podemos escribirlo como:

$$\log(Y_t) = \beta_0 + \beta_1 t$$

Este modelo es lineal, pero la diferencia es que la variable dependiente tiene aplicado el logaritmo. Estos modelos se conocen como semilog porque solo una variable aparece en forma logaritmica. En este modelo, el coeficiente de la pendiente mide el cambio proporcional constante relativo en Y para un cambio absoluto en el valor de la regresora:

$$\beta_1 = \frac{\Delta\%Y}{\Delta X}$$

Esto es un cambio porcentual o tasa de crecimiento en Y ocasionada por un cambio absoluto en X (en algunos libros se conoce como la semielasticidad)

6. Modelo Lin-Log (Rendimientos decrecientes)

A diferencia del modelo pasado, en este caso la variable que tiene logaritmo es la variable independiente:

Sea el modelo: $Y_i = \beta_0 + \beta_1 \log(X_i)$

Este modelo se conoce como **modelo lin-log**.

Cuando se transforma el modelo de esta manera, la correspondiente β_1 mide el cambio absoluto de Y vs el cambio porcentual de X , es decir: $\frac{\Delta Y}{\Delta \% X}$

7. Resumen

```
knitr::kable(tibble(  
  Modelo = c("Lineal", "Lineal estandarizado", "log-log", "lin-log", "log-lin"),  
  `Si x aumenta` = c("1 unidad", "1 sd", "1%", "1%", "1 unidad"),  
  `Entonces y incrementa` = c("b1 unidades", "b1 sd", "b1 %", "b1 unidades", "b1%")  
))
```

Modelo	Si x aumenta	Entonces y incrementa
Lineal	1 unidad	b1 unidades
Lineal estandarizado	1 sd	b1 sd
log-log	1%	b1 %
lin-log	1%	b1 unidades
log-lin	1 unidad	b1%