Análisis de Regresión

Dora Suárez, Juan F. Pérez

Departamento MACC
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario

juanferna.perez@urosario.edu.co

Primer Semestre de 2019

Contenidos

Regresión

Regresión

- Representar la posible asociación entre dos o más variables

Regresión

- Representar la posible asociación entre dos o más variables
- Regresión simple: dos variables (X y Y)

Regresión

- Representar la posible asociación entre dos o más variables
- Regresión simple: dos variables (X y Y)
- X: variable independiente o regresor

Regresión

- Representar la posible asociación entre dos o más variables
- Regresión simple: dos variables (X y Y)
- X: variable independiente o regresor
- Y: variable dependiente o respuesta

Regresión

- Representar la posible asociación entre dos o más variables
- Regresión simple: dos variables (X y Y)
- X: variable independiente o regresor
- Y: variable dependiente o respuesta
- $$Y \approx f(X)$$

Regresión - Ejemplos

- Ventas Y y publicidad X

Regresión - Ejemplos

- Ventas Y y publicidad X
- Beneficios Y e inversión X

Regresión - Ejemplos

- Ventas Y y publicidad X
- Beneficios Y e inversión X
- Resultados de pruebas Y e inversión en infraestructura X

Regresión - Ejemplos

- Ventas Y y publicidad X
- Beneficios Y e inversión X
- Resultados de pruebas Y e inversión en infraestructura X
- Resultados de pruebas Y e inversión en formación X

Regresión - Ejemplos

- Ventas Y y publicidad X
- Beneficios Y e inversión X
- Resultados de pruebas Y e inversión en infraestructura X
- Resultados de pruebas Y e inversión en formación X
- $$Y \approx f(X)$$

Regresión - Preguntas

- ¿Existe un relación entre las dos variables?

Regresión - Preguntas

- ¿Existe un relación entre las dos variables?
- ¿Qué tal fuerte es la relación?

Regresión - Preguntas

- ¿Existe un relación entre las dos variables?
- ¿Qué tal fuerte es la relación?
- ¿Qué tan bien (precisamente) se puede estimar el efecto de X en Y?

Regresión - Preguntas

- ¿Existe un relación entre las dos variables?
- ¿Qué tal fuerte es la relación?
- ¿Qué tan bien (precisamente) se puede estimar el efecto de X en Y?
- ¿Qué tan bien (precisamente) se puede pronosticar el valor de Y dado un valor de X?

Regresión - Preguntas

- ¿Existe un relación entre las dos variables?
- ¿Qué tal fuerte es la relación?
- ¿Qué tan bien (precisamente) se puede estimar el efecto de X en Y?
- ¿Qué tan bien (precisamente) se puede pronosticar el valor de Y dado un valor de X?
- ¿Cómo es la relación? ¿Lineal? ¿No lineal?

$$Y \approx f(X)$$

Regresión Lineal Simple

Regresión Lineal Simple

- $$Y \approx \beta_0 + \beta_1 X$$

Regresión Lineal Simple

- $$Y \approx \beta_0 + \beta_1 X$$
- β_0: valor de Y cuando $X = 0$ (intercepto)

Regresión Lineal Simple

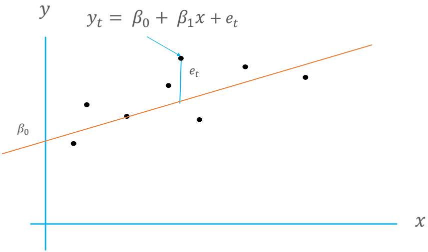
- $$Y \approx \beta_0 + \beta_1 X$$
- β_0: valor de Y cuando $X = 0$ (intercepto)
- β_1: efecto de X en Y (pendiente)

Regresión Lineal Simple

- $$Y \approx \beta_0 + \beta_1 X$$
- β_0: valor de Y cuando $X = 0$ (intercepto)
- β_1: efecto de X en Y (pendiente)
- ¿Cuánto cambia el valor de Y cuando X cambia su valor en 1?

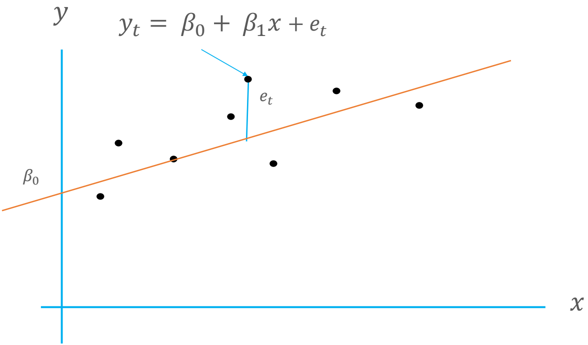
Regresión Lineal Simple

- $$Y \approx \beta_0 + \beta_1 X$$
- β_0: valor de Y cuando $X = 0$ (intercepto)
- β_1: efecto de X en Y (pendiente)
- ¿Cuánto cambia el valor de Y cuando X cambia su valor en 1?



Análisis descriptivo de datos (ejemplo)

$$Y = \beta_0 + \beta_1 X + e$$



Regresión Lineal Simple

- $$Y = \beta_0 + \beta_1 X + e$$

Regresión Lineal Simple

- $$Y = \beta_0 + \beta_1 X + e$$
- Observaciones: (y_t, x_t) para $t = 1, \ldots, n$

Regresión Lineal Simple

- $$Y = \beta_0 + \beta_1 X + e$$
- Observaciones: (y_t, x_t) para $t = 1, \ldots, n$
- Modelo: $Y = \beta_0 + \beta_1 X$

Regresión Lineal Simple

- $$Y = \beta_0 + \beta_1 X + e$$
- Observaciones: (y_t, x_t) para $t = 1, \ldots, n$
- Modelo: $Y = \beta_0 + \beta_1 X$
- Estimar β_0 y β_1

Regresión Lineal Simple

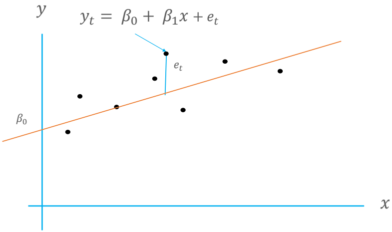
- $$Y = \beta_0 + \beta_1 X + e$$
- Observaciones: (y_t, x_t) para $t = 1, \ldots, n$
- Modelo: $Y = \beta_0 + \beta_1 X$
- Estimar β_0 y β_1
- Estimadores: $\hat{\beta}_0$ y $\hat{\beta}_1$

Regresión Lineal Simple

- $$Y = \beta_0 + \beta_1 X + e$$
- Observaciones: (y_t, x_t) para $t = 1, \ldots, n$
- Modelo: $Y = \beta_0 + \beta_1 X$
- Estimar β_0 y β_1
- Estimadores: $\hat{\beta}_0$ y $\hat{\beta}_1$
- Valor estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

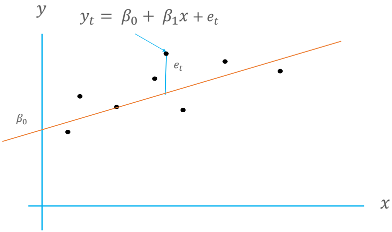
Regresión Lineal Simple

- Valor real: y_t



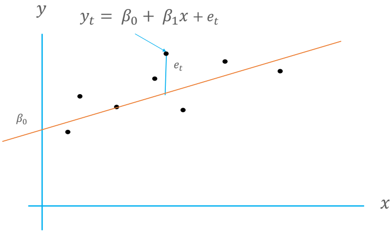
Regresión Lineal Simple

- Valor real: y_t
- Valor estimado: $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x$



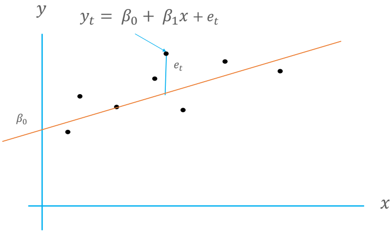
Regresión Lineal Simple

- Valor real: y_t
- Valor estimado: $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x$
- Error: $e_t = y_t - \hat{y}_t$



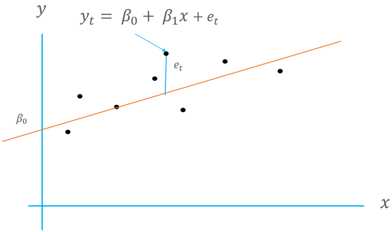
Regresión Lineal Simple

- Valor real: y_t
- Valor estimado: $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x$
- Error: $e_t = y_t - \hat{y}_t$
- Error de estimación / residual



Regresión Lineal Simple

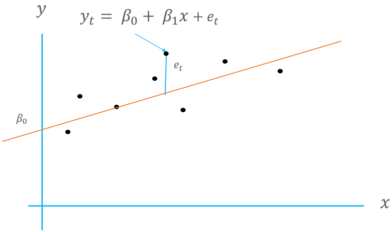
- Error: $e_t = y_t - \hat{y}_t = y - t - \hat{\beta}_0 - \hat{\beta}_1 x_t$



Regresión Lineal Simple

- Error: $e_t = y_t - \hat{y}_t = y - t - \hat{\beta}_0 - \hat{\beta}_1 x_t$
- Suma de los errores/residuales al cuadrado (RSS):

$$RSS = e_1^2 + \cdots + e_n^2 = \sum_{t=1}^{n} e_t^2$$



Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_t)^2$$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_t)^2$$

- Minimizar RSS: mejores valores de $\hat{\beta}_0$ y $\hat{\beta}_1$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_t)^2$$

- Minimizar RSS: mejores valores de $\hat{\beta}_0$ y $\hat{\beta}_1$
-

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_t)^2$$

- Minimizar RSS: mejores valores de $\hat{\beta}_0$ y $\hat{\beta}_1$
-
$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$
-
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 + \hat{\beta}_1 x_t)^2$$

- Minimizar RSS: mejores valores de $\hat{\beta}_0$ y $\hat{\beta}_1$
-
$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$
-
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
- $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$, $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$

Método de mínimos cuadrados

- Estimadores de mínimos cuadrados:

Método de mínimos cuadrados

- Estimadores de mínimos cuadrados:
-
$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

Método de mínimos cuadrados

- Estimadores de mínimos cuadrados:
- $$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$
- $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Método de mínimos cuadrados

- Estimadores de mínimos cuadrados:
- $$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$
- $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
- Valor estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Regresión Simple en R

```
data("iris")
View(iris)

names(iris)
modelo <- lm(iris$Sepal.Length~iris$Sepal.Width)

modelo <- lm(Sepal.Length~Sepal.Width, data = iris)


attach(iris)
modelo <- lm(Sepal.Length~Sepal.Width)

summary(modelo)
```

Regresión Simple en R

```
attach(iris)
modelo <- lm(Sepal.Length~Sepal.Width)
summary(modelo)

library(ggplot2)

ggplot(iris, aes(x=Petal.Width, y=Sepal.Width)) +
  geom_point() +
  stat_smooth(method="lm", col = "red")
```

Regresión Simple en R

```
attach(iris)
modelo2 <- lm(Sepal.Length~Petal.Width)
summary(modelo2)

ggplot(iris, aes(x=Petal.Width, y=Sepal.Length)) +
  geom_point() +
  stat_smooth(method="lm", col = "red")
```

Método de mínimos cuadrados

- Error estándar de los estimadores:

Método de mínimos cuadrados

- Error estándar de los estimadores:
- $$SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{t=1^n} (x_t - \bar{x})^2} \right)$$

Método de mínimos cuadrados

- Error estándar de los estimadores:
- $$SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{t=1^n} (x_t - \bar{x})^2} \right)$$
- $$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1^n} (x_t - \bar{x})^2}$$

Método de mínimos cuadrados

- Error estándar de los estimadores:
- $$SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{t=1^n} (x_t - \bar{x})^2} \right)$$
- $$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1^n} (x_t - \bar{x})^2}$$
- e_t no correlacionados con media 0 y varianza σ^2

Método de mínimos cuadrados

- Error estándar de los estimadores:
- $$SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{t=1^n} (x_t - \bar{x})^2} \right)$$
- $$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1^n} (x_t - \bar{x})^2}$$
- e_t no correlacionados con media 0 y varianza σ^2
- Estimador de σ^2
$$\hat{\sigma}^2 = \sqrt{RSS/(n-2)}$$

Método de mínimos cuadrados

- Error estándar de los estimadores:
- $$SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{t=1^n} (x_t - \bar{x})^2} \right)$$
- $$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1^n} (x_t - \bar{x})^2}$$
- e_t no correlacionados con media 0 y varianza σ^2
- Estimador de σ^2
$$\hat{\sigma}^2 = \sqrt{RSS/(n-2)}$$
- RSS$=\sum_{t=1}^{n} e_t$

R^2

- Medida de bondad del modelo

R^2

- Medida de bondad del modelo
- Suma total de cuadrados

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

R^2

- Medida de bondad del modelo
- Suma total de cuadrados

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

- Suma de cuadrados del error/residuales

$$RSS = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

R^2

- Medida de bondad del modelo
- Suma total de cuadrados

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

- Suma de cuadrados del error/residuales

$$RSS = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

- $$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2

- Medida de bondad del modelo
- Suma total de cuadrados

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

- Suma de cuadrados del error/residuales

$$RSS = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

-

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Propoción de la varianza de Y explicada por X (el modelo)

R^2

- ¿Cuál es un buen valor de R^2?

R^2

- ¿Cuál es un buen valor de R^2?
- Correlación:

$$\rho(X, Y) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

R^2

- ¿Cuál es un buen valor de R^2?
- Correlación:

$$\rho(X,Y) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

- En el caso de la regresión lineal simple

$$R^2 = \rho(X,Y)^2$$

Regresión Lineal Múltiple

Regresión Lineal Múltiple

- Múltiple variables independientes o regresores

Regresión Lineal Múltiple

- Múltiple variables independientes o regresores
- Explicar/predecir la variable dependiente o respuesta

Regresión Lineal Múltiple

- Múltiple variables independientes o regresores
- Explicar/predecir la variable dependiente o respuesta
- $$Y = f(X_1, X_2, \ldots, X_K)$$

Regresión Lineal Múltiple

- Múltiple variables independientes o regresores
- Explicar/predecir la variable dependiente o respuesta
- $$Y = f(X_1, X_2, \ldots, X_K)$$
- Modelo lineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + e$$

Regresión Lineal Múltiple

- Múltiple variables independientes o regresores
- Explicar/predecir la variable dependiente o respuesta
- $$Y = f(X_1, X_2, \ldots, X_K)$$
- Modelo lineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + e$$

- β_j: valor esperado del cambio en Y dado un incremento de una unidad en X_j

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \quad = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_K x_{Kt}$$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \quad = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_K x_{Kt}$$

- Minimizar RSS: mejores valores de $\hat{\beta}_j$

Método de mínimos cuadrados

- Suma de los residuales al cuadrado (RSS):

$$RSS = \sum_{t=1}^{n} e_t^2 = \quad = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_K x_{Kt}$$

- Minimizar RSS: mejores valores de $\hat{\beta}_j$
-

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Evaluando el modelo

- ¿Es al menos uno de los regresores $X_1, \dots, X_K$ útil para predecir el valor de Y?

Evaluando el modelo

- ¿Es al menos uno de los regresores $X_1, \ldots, X_K$ útil para predecir el valor de Y?
- H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$
- H_a: al menos un $\beta_j \neq 0$

Evaluando el modelo

- ¿Es al menos uno de los regresores $X_1, \ldots, X_K$ útil para predecir el valor de Y?
- H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$
- H_a: al menos un $\beta_j \neq 0$
- Estadístico F

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

Evaluando el modelo

- ¿Es al menos uno de los regresores $X_1, \ldots, X_K$ útil para predecir el valor de Y?
- H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$
- H_a: al menos un $\beta_j \neq 0$
- Estadístico F
$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$
- Si H_0 es cierta, F es cercano a 1

Evaluando el modelo

- ¿Es al menos uno de los regresores $X_1, \ldots, X_K$ útil para predecir el valor de Y?
- H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$
- H_a: al menos un $\beta_j \neq 0$
- Estadístico F
$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$
- Si H_0 es cierta, F es cercano a 1
- Si H_a es cierta, F es mayor a 1

Evaluando el modelo

- Estadístico F

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

- Si H_0 es cierta, F es cercano a 1
- Si H_a es cierta, F es mayor a 1

Evaluando el modelo

- Estadístico F

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

- Si H_0 es cierta, F es cercano a 1
- Si H_a es cierta, F es mayor a 1
- Si los errores se distribuyen normal, F sigue una distribución F y calculamos un valor p (decidir sobre el rechazo o no de H_0)

Selección de variables

- Probar varios modelos y seleccionar el mejor

Selección de variables

- Probar varios modelos y seleccionar el mejor
- Criterio de decisión

Selección de variables

- Probar varios modelos y seleccionar el mejor
- Criterio de decisión
- AIC: Akaike Information Criterion

Selección de variables

- Probar varios modelos y seleccionar el mejor
- Criterio de decisión
- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion

Selección de variables

- Probar varios modelos y seleccionar el mejor
- Criterio de decisión
- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion
- R^2 ajustado

Selección de variables

Selección hacia adelante:

- Empezar con β_0 y buscar la mejor variable a agregar

Selección de variables

Selección hacia adelante:

- Empezar con β_0 y buscar la mejor variable a agregar
- Buscar la segunda mejor variable a agregar

Selección de variables

Selección hacia adelante:

- Empezar con β_0 y buscar la mejor variable a agregar
- Buscar la segunda mejor variable a agregar
- Continuar hasta que no valga la pena agregar más variables (criterio de parada)

Selección de variables

Selección hacia atrás:

- Empezar con todas las variables

Selección de variables

Selección hacia atrás:

- Empezar con todas las variables
- Decartar la menos relevante para el modelo (e.g., la de mayor valor p)

Selección de variables

Selección hacia atrás:

- Empezar con todas las variables
- Decartar la menos relevante para el modelo (e.g., la de mayor valor p)
- Continuar hasta que ninguna variable sea candidata a salir

Selección de variables

Selección mixta:

- Empezar como en selección adelante y agregar variables

Selección de variables

Selección mixta:

- Empezar como en selección adelante y agregar variables
- Si al agregar una variable se incrementa el valor p de otra por encima de un umbral, se descarta esta última variable

Selección de variables

Selección mixta:

- Empezar como en selección adelante y agregar variables
- Si al agregar una variable se incrementa el valor p de otra por encima de un umbral, se descarta esta última variable
- Continuar hasta que no haya variables con potencial de entrar ni variables candidatas a salir

Regresión Lineal Múltiple en R

```
library(MASS)
View(Boston)
names(Boston)
?Boston

attach(Boston)
modelo <- lm(medv~indus+crim)
summary(modelo)
```

Regresión Lineal Múltiple en R

```
modelo <- lm(medv~.,data=Boston)
summary(modelo)

modelo <- lm(medv~.-age,data=Boston)
summary(modelo)

modelo <- lm(medv~.-age-indus,data=Boston)
summary(modelo)

modelo <- lm(medv~crim*zn,data=Boston)
summary(modelo)
```

Regresión Lineal Múltiple en R

```
modelo <- lm(medv~lstat+I(lstat^2),data=Boston)
summary(modelo)

modelo <- lm(medv~poly(lstat,2),data=Boston)
summary(modelo)

modelo <- lm(medv~poly(lstat,5),data=Boston)
summary(modelo)

modelo <- lm(medv~poly(lstat,7),data=Boston)
summary(modelo)
```

Regresión Lineal Múltiple en R

```
library(ggplot2)
modelo <- lm(medv ~ poly(lstat,5), data = Boston)
pred <- data.frame(
   lstat = seq(
            from = range(Boston$lstat)[1],
      to = range(Boston$lstat)[2],
                  length.out = nrow(Boston))
)
err <- predict(modelo, newdata = pred, se.fit = TRUE)
```

Regresión Lineal Múltiple en R

```
pred$lci <- err$fit - 1.96 * err$se.fit
pred$fit <- err$fit
pred$uci <- err$fit + 1.96 * err$se.fit

ggplot(pred, aes(x = lstat, y = fit)) +
  theme_bw() +
  geom_line() +
  geom_smooth(aes(ymin = lci, ymax = uci),
                              stat = "identity") +
  geom_point(data = Boston, aes(x = lstat, y = medv))
```

Regresión Lineal Múltiple en R

```
rho <- cor(Boston)
rho["medv",]
pairs(Boston[,c("medv","rm","lstat")])

modelo <- lm(medv~rm+lstat,data=Boston)
summary(modelo)

modelo <- lm(medv~poly(rm,2)+poly(lstat,2),
             data=Boston)
summary(modelo)
```

Interacciones

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$

Interacciones

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$

Interacciones

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$
- $Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + e$

Interacciones

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$
- $Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + e$
- $Y = \beta_0 + \beta_1 (X_2) X_1 + \beta_2 X_2 + e$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ 0, & \text{si la persona es hombre} \end{cases}$$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ 0, & \text{si la persona es hombre} \end{cases}$$
- $Y_t = \beta_0 + \beta_1 X_t + e_t$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ 0, & \text{si la persona es hombre} \end{cases}$$
- $Y_t = \beta_0 + \beta_1 X_t + e_t$
- $$Y_t = \begin{cases} \beta_0 + \beta_1 + e_t, & \text{si la persona es mujer,} \\ \beta_0 + e_t, & \text{si la persona es hombre} \end{cases}$$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ -1, & \text{si la persona es hombre} \end{cases}$$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ -1, & \text{si la persona es hombre} \end{cases}$$
- $Y_t = \beta_0 + \beta_1 X_t + e_t$

Variables Cualitativas

Codificación y modelo:

- $$X_t = \begin{cases} 1, & \text{si la persona es mujer,} \\ -1, & \text{si la persona es hombre} \end{cases}$$
- $Y_t = \beta_0 + \beta_1 X_t + e_t$
- $$Y_t = \begin{cases} \beta_0 + \beta_1 + e_t, & \text{si la persona es mujer,} \\ \beta_0 - \beta_1 + e_t, & \text{si la persona es hombre} \end{cases}$$

Variables Cualitativas

Codificación y modelo (más de dos categorías):

- $$X_{1t} = \begin{cases} 1, & \text{ubicación ideal,} \\ 0, & \text{ubicación no ideal} \end{cases}$$

Variables Cualitativas

Codificación y modelo (más de dos categorías):

- $$X_{1t} = \begin{cases} 1, & \text{ubicación ideal,} \\ 0, & \text{ubicación no ideal} \end{cases}$$

- $$X_{2t} = \begin{cases} 1, & \text{ubicación promedio,} \\ 0, & \text{ubicación no promedio} \end{cases}$$

Variables Cualitativas

Codificación y modelo (más de dos categorías):

- $$X_{1t} = \begin{cases} 1, & \text{ubicación ideal,} \\ 0, & \text{ubicación no ideal} \end{cases}$$
- $$X_{2t} = \begin{cases} 1, & \text{ubicación promedio,} \\ 0, & \text{ubicación no promedio} \end{cases}$$
- Si $X_{1t} = X_{2t} = 0$, ubicación mala

Variables Cualitativas

Codificación y modelo (más de dos categorías):

- $$X_{1t} = \begin{cases} 1, & \text{ubicación ideal,} \\ 0, & \text{ubicación no ideal} \end{cases}$$
- $$X_{2t} = \begin{cases} 1, & \text{ubicación promedio,} \\ 0, & \text{ubicación no promedio} \end{cases}$$
- Si $X_{1t} = X_{2t} = 0$, ubicación mala
- $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + e_t$

Variables Cualitativas

Codificación y modelo (más de dos categorías):

- $$X_{1t} = \begin{cases} 1, & \text{ubicación ideal,} \\ 0, & \text{ubicación no ideal} \end{cases}$$
- $$X_{2t} = \begin{cases} 1, & \text{ubicación promedio,} \\ 0, & \text{ubicación no promedio} \end{cases}$$
- Si $X_{1t} = X_{2t} = 0$, ubicación mala
- $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + e_t$
- $$Y_t = \begin{cases} \beta_0 + \beta_1 + e_t, & \text{ubicación ideal,} \\ \beta_0 + \beta_2 + e_t, & \text{ubicación promedio,} \\ \beta_0 + e_t, & \text{ubicación mala} \end{cases}$$

Variables Cualitativas

```
library(ISLR)
View(Carseats)
names(Carseats)

lm.fit =lm(Sales~.
        +Income*Advertising
        +Price*Age,
        data=Carseats )
summary (lm.fit)

contrasts(Carseats$ShelveLoc)
```