Probabilidad y Estadística 1

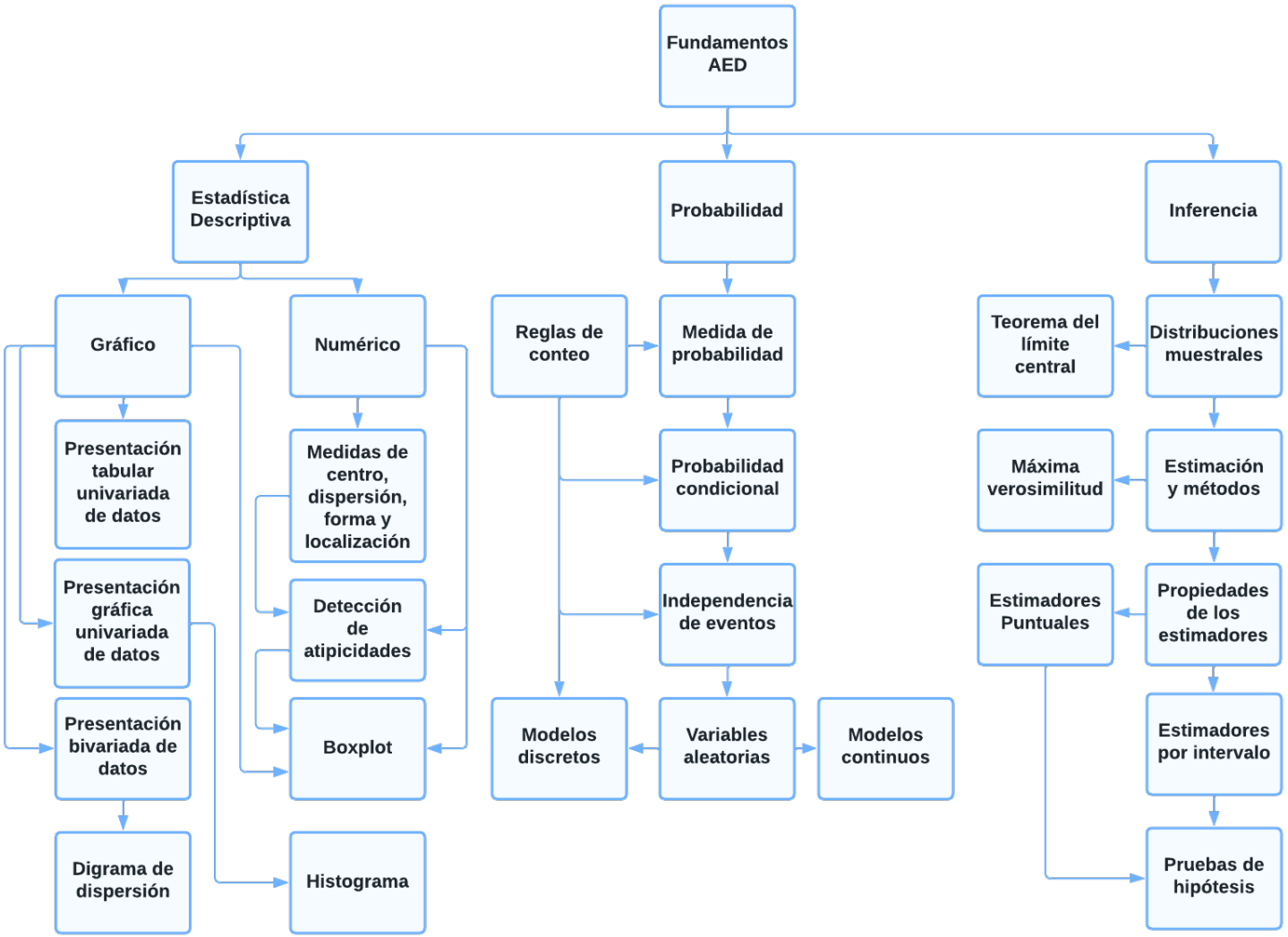
Modelos lineales

Profesor: Nicolás López MSc

Semana 15

Estadística introductoria para el modelo de RLS

Estadística introductoria para el modelo de RLS



Aprendizaje estadístico

Introducción

Aprender el proceso subyacente generador de datos. El proceso es formalizado matemáticamente en el aprendizaje estadístico y se clasifica en dos grupos:

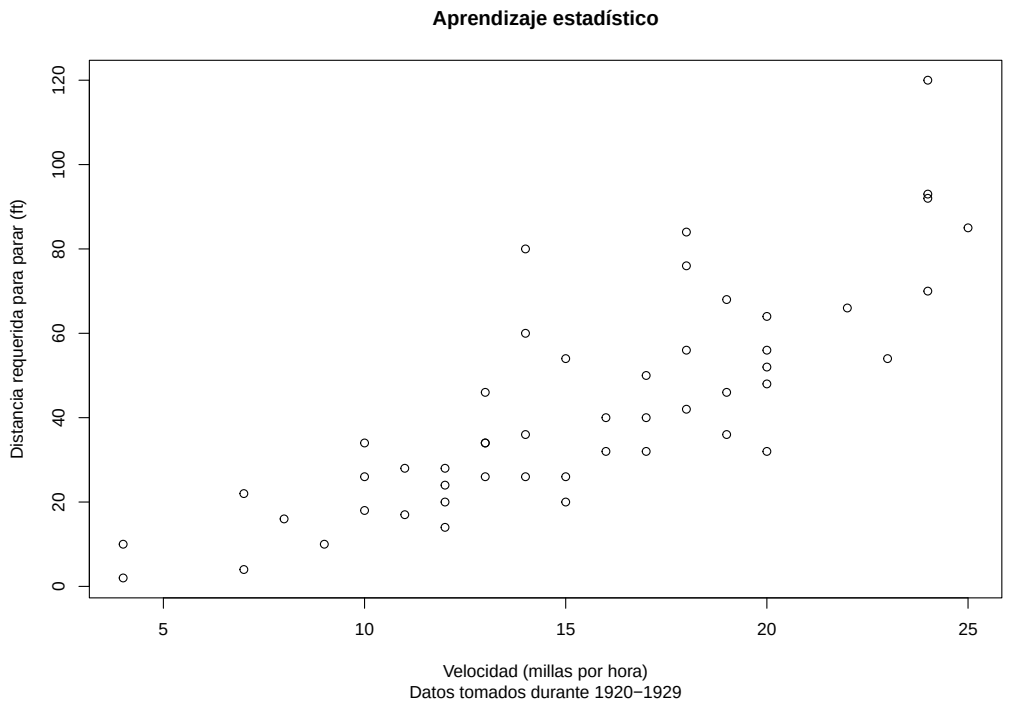
- Aprendizaje supervisado: Se tiene un resultado (*outcome*) que guía el proceso de aprendizaje (ej. identificación de dígitos).
- Aprendizaje no supervisado. No se tiene una medición de un resultado para guiar el aprendizaje (ej. clasificación de carros basado en características).

En ambos escenarios se cuenta con un conjunto de covariables (*features*) que permiten el aprendizaje.

¿Aprendizaje supervisado o no supervisado?

Ejemplo 1

```
plot(cars, xlab = "Velocidad (millas por hora)",
           ylab = "Distancia requerida para parar (ft)",
           main  = "Aprendizaje estadístico",
           sub   = "Datos tomados durante 1920-1929")
```



Ejemplo 2

```
#mnist = dslabs::read_mnist()
#i  = 10
#image(1:28, 1:28, matrix(mnist$test$images[i,], nrow=28)[ , 28:1],
#                   col  = gray(seq(0, 1, 0.05)),
#                   xlab = paste0("Número ",mnist$test$labels[i]),
#                   ylab = "", main = "Aprendizaje estadístico")
```

Tipos de variables

En el aprendizaje estadístico contamos con dos clases principales de variables:

- Cuantitativas
- Cualitativas

Esto tanto para las covariables como para la variable respuesta. Existe un mayor refinamiento en la categorización, pero por ahora basta entender que **el mismo problema de aprendizaje (supervisado o no) puede darse para diferente naturaleza de las variables**:

- Análisis de regresión lineal (simple/múltiple): Supervisado con respuesta cuantitativa.

Definiciones importantes

Se destacan 4 elementos fundamentales en el aprendizaje estadístico dada la revisión anterior:

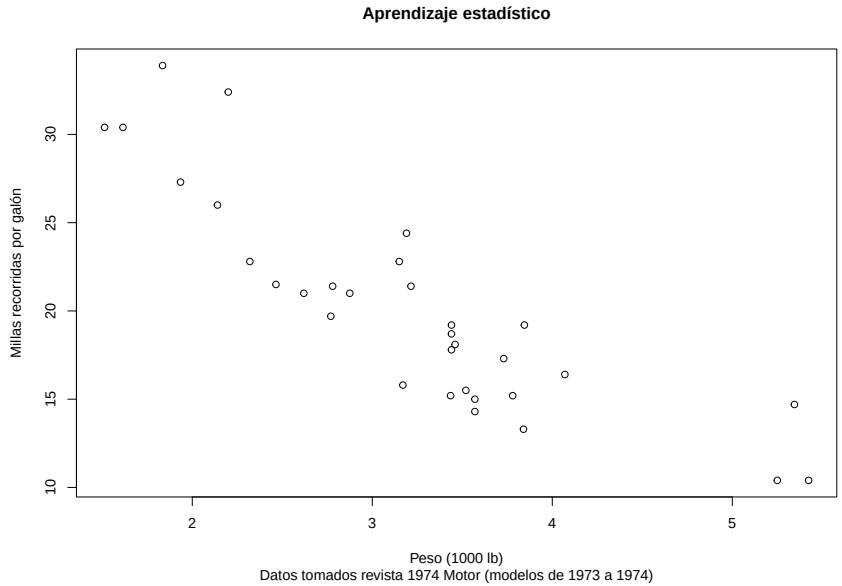
- Proceso generador P.
- Variable de entrada/covariable/input: X (uni/multivariada).
- Variable de salida/variable respuesta/output: Y (univariada usualmente).
- Observaciones/realizaciones/mediciones: $(x_1, y_1), ..., (x_n, y_n)$.

Estas mediciones son arregladas en una matriz dise~no $\mathbf{X}$.

RLS

S14. Introducción

Si analizamos con detenimiento la gráfica de dispersión de los datos de velocidad podemos establecer con claridad una relación entre estas dos variables.



Si Y representa 'Millas recorridas por galón' y x es igual a 'Peso (1000 lb)', podemos representar una relación **determinística** entre las variables de la siguiente forma:

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

Por lo tanto una relación **aleatoria** que ajusta por el error está formulado por:

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

La ecuación anterior se llama *Modelo de Regresión Lineal Simple*.

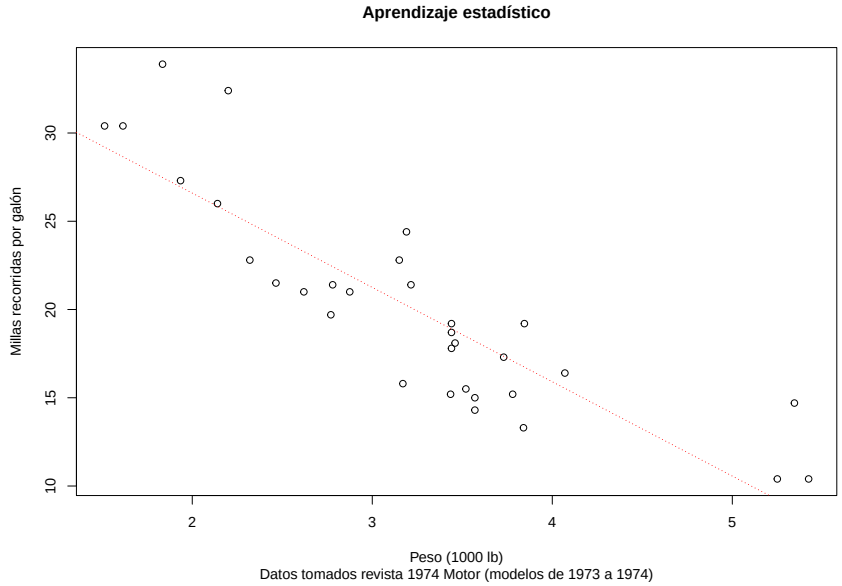
El modelo aleatorio esta completamente especificado cuando definimos las características aleatorias del error, suponemos que $\epsilon \sim N(0, \sigma^2)$. Tanto el error como la respuesta son aleatorias, bajo normalidad de ϵ, Y también es normal (¿por qué?). Entonces la respuesta esperada de Y dado X es:

$$E(Y|X = x) = \mu_{Y|X=x} = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$$

Con una varianza igual a

$$V(Y|X = x) = \sigma_{Y|X=x} = V(\beta_0 + \beta_1 x + \epsilon) = V(\epsilon) = \sigma^2$$

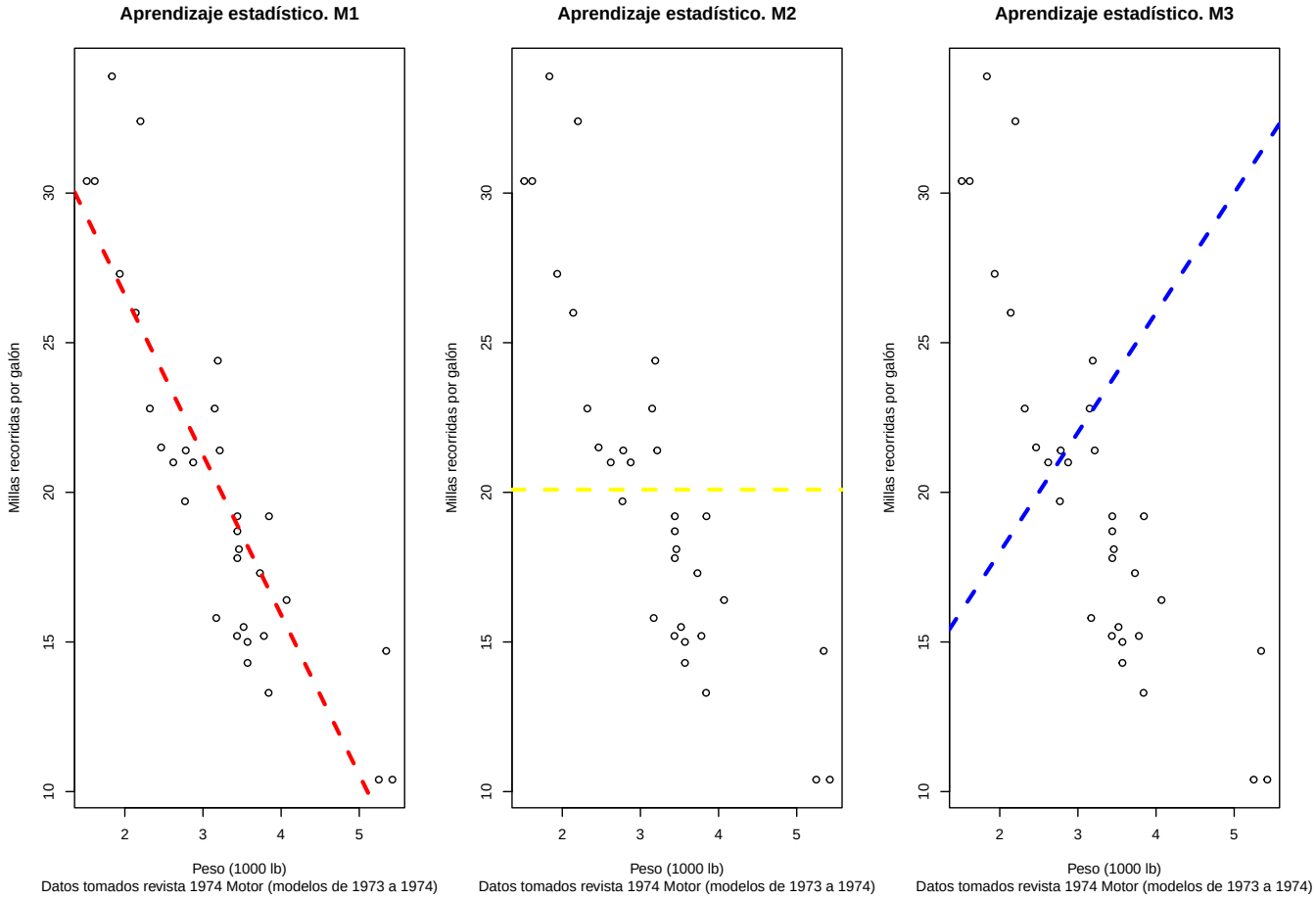
El modelo de regresión $\mu_{Y|X}$ es una línea recta de valores promedios, esto es, la altura de la línea de regresión en X es el valor esperado $\beta_0 + \beta_1 X$. Esto implica que hay una distribución de valores de Y para cada X, y que la varianza σ^2 de esta distribución es igual en cada X.



Con una realización de n observaciones $(x_1, y_1), ..., (x_n, y_n)$ se estiman los parámetros del modelo: $\hat{\beta}_0$, $\hat{\beta}_1$, y $\hat{\sigma}$.

Ajuste por mínimos cuadrados

Para encontrar la línea que se ajusta mejor a los datos, necesitamos una medida de calidad del ajuste. Bajo estos tres candidatos es claro cual resulta en una menor **suma de cuadrados**:



Es claro que:

$$SC(M1) < SC(M2) < SC(M3)$$

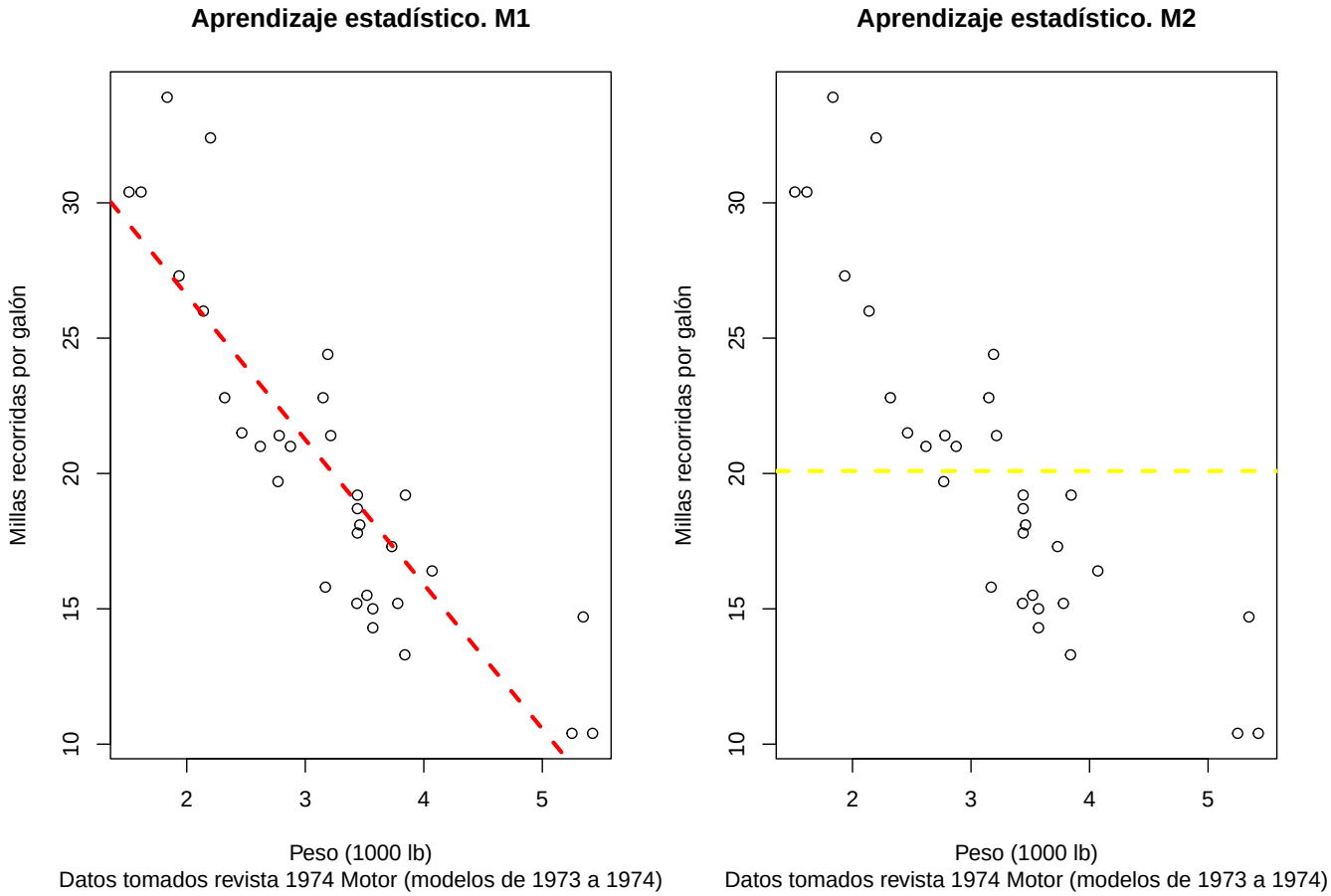
Con

$$SC(Mj) = \sum_i (y_i - \hat{y}_{i,Mj})^2$$

En dónde

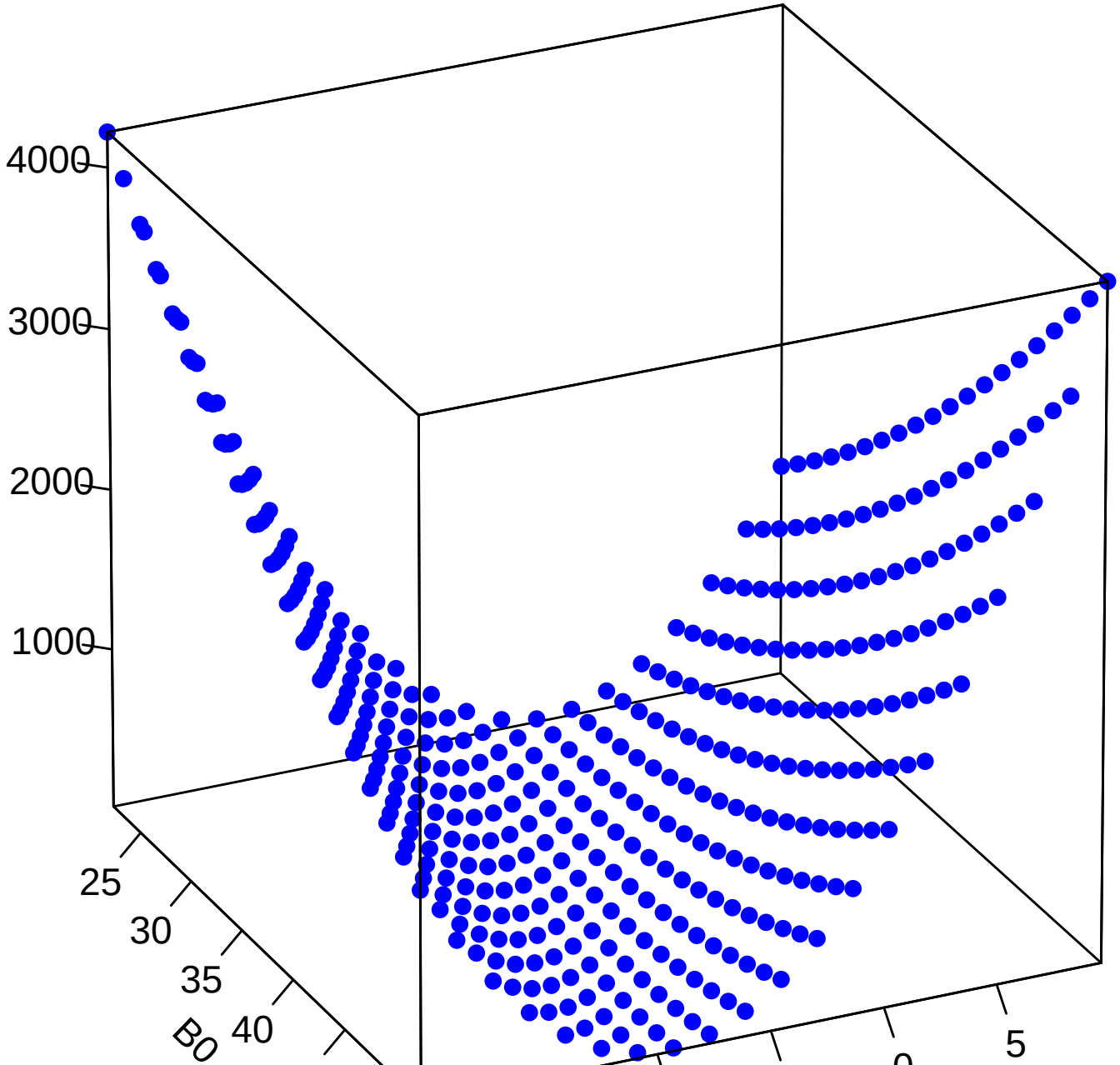
- El **modelo simple o reducido** (M2) no utiliza información de X para encontrar el valor de Y, asume el valor promedio de Y como modelo marginal. Este es considerado la línea base (*baseline*).
- La estimación por mínimos cuadrados consiste en encontrar β_0 y β_1 de tal forma que minimicen la suma de cuadrados, es decir, los **residuales cuadrados**. (¿qué sucedió con σ en la estimación?).

Es evidente que hay menor variabilidad alrededor de $M1$ que alrededor de $M2$, es decir que la variación de las millas recorridas es explicada por el peso del vehículo. ¿Cómo formalizar esta noción?



La SC del modelo simple ($SC(M2) = SC_T$) cuantifica la variabilidad total de Y respecto a su media. Por su parte $SC(M1) = SC_E$ mide la variación remanente al ajustar el modelo mediante mínimos cuadrados.

Fíjese que para cada modelo se tiene un intercepto y una pendiente (β_0, β_1), y con ello se obtiene SC_E, es decir $SC_E(\beta_0, \beta_1)$. Los valores estimados $(\hat{\beta}_0, \hat{\beta}_1)$ por mínimos cuadrados minimizan la función de error:



Para cualquier pareja de parámetros, comparar SC_T con SC_E cuantifica la reducción de la variabilidad bajo el modelo lineal en X. La reducción de la varianza en Y explicada por X bajo el modelo es igual a:

$$SC_M = SC_T - SC_E$$

Así SC_M cuantifica la reducción en la variación total al ajustar el modelo lineal en X. Al igual que SC_E, SC_M es función de (β_0, β_1), es decir $SC_M(\beta_0, \beta_1)$, la cual es maximizada en $(\hat{\beta}_0, \hat{\beta}_1)$ por mínimos cuadrados (¿qué unidades tiene SC_M?).

La SC_M es estandarizada como:

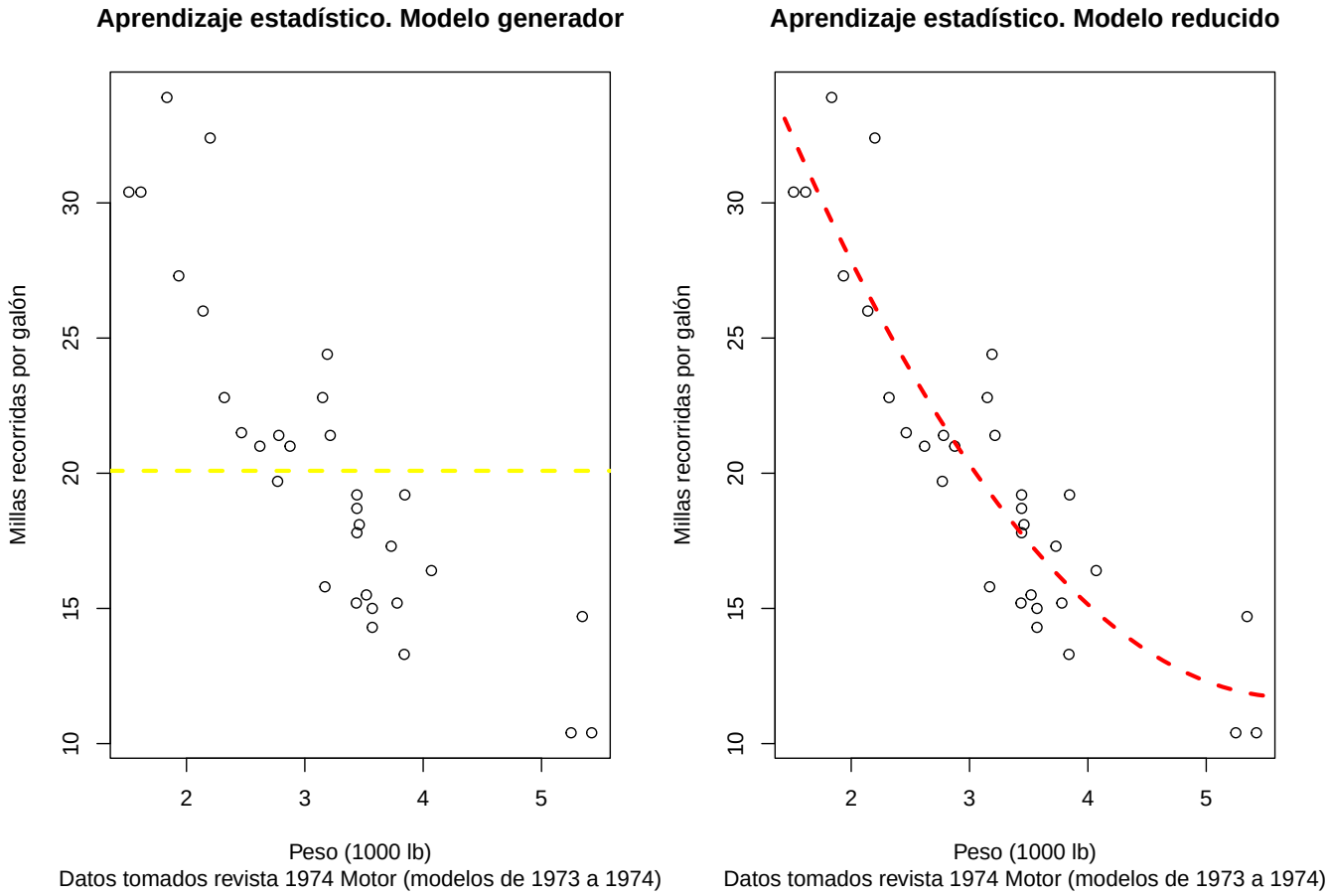
$$R^2 = \frac{SC_M}{SC_T}$$

- Con R^2 cuantificamos la **proporción** de la varianza en Y explicada por el regresor x.
- Al ser R^2 cercano a 1, SC_M se acerca a SC_T, es decir que el *modelo explica* la variabilidad en Y (¿cómo lo medimos objetivamente?).
- Al ser R^2 cercano a 0, SC_M se aleja de SC_T, es decir que el *modelo no explica* la variabilidad en Y (¿cómo lo medimos objetivamente?).

En nuestro ejemplo $R^2 = 0.75$, con lo cual hay una reducción de la varianza de un 75% en las millas recorridas al considerar linealmente el peso del vehículo.

Note que esta definición de R^2 aplica para situaciones aún más generales, por ejemplo, para un modelo con componente lineal y **cuadrático** en X:

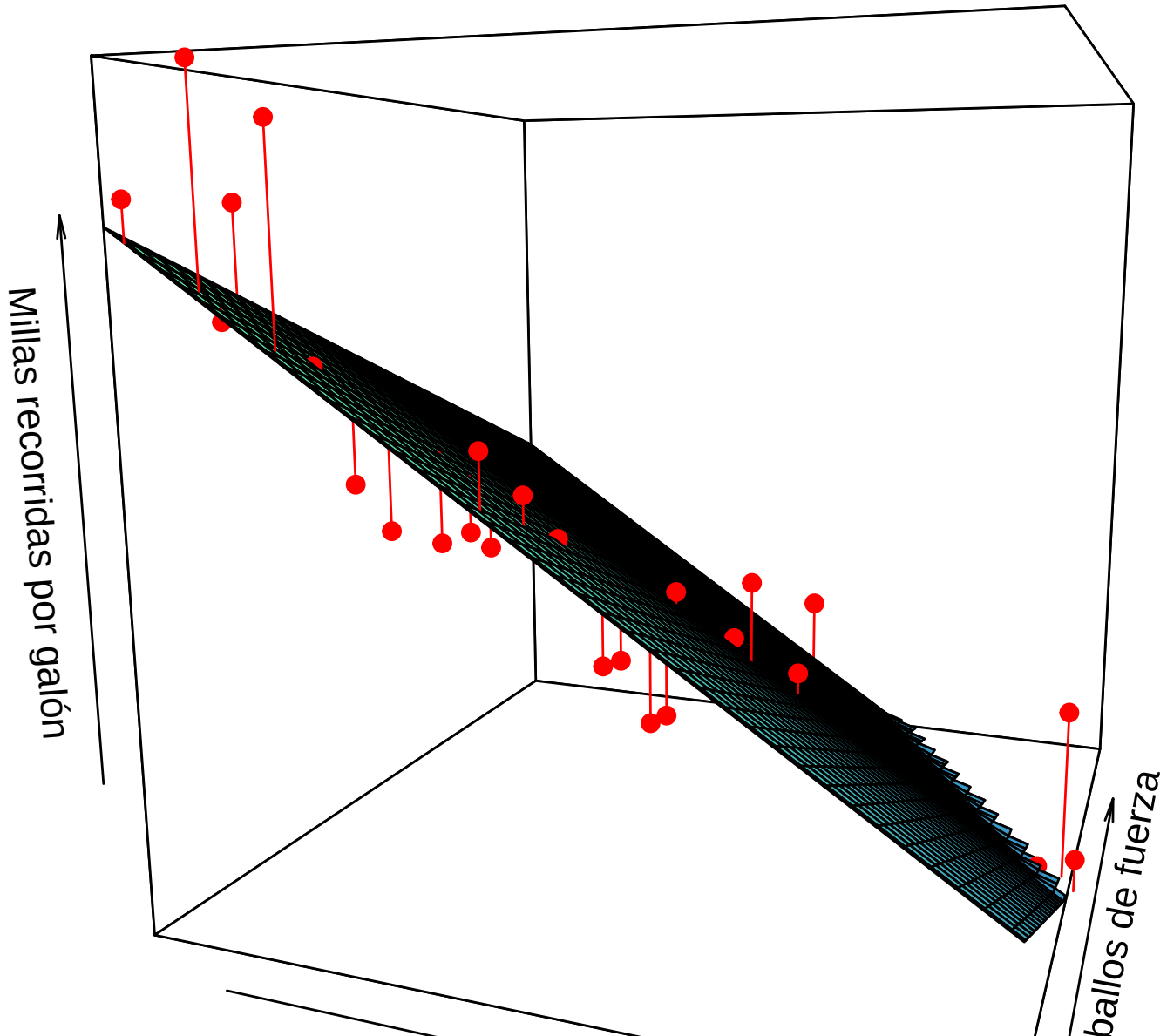
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$



Y bajo este modelo, tenemos un $R^2 = 0.81$.

También para el escenario multivariado, con z igual a los caballos de fuerza del vehículo

$$Y = \beta_0 + \beta_1 x + \beta_2 z$$



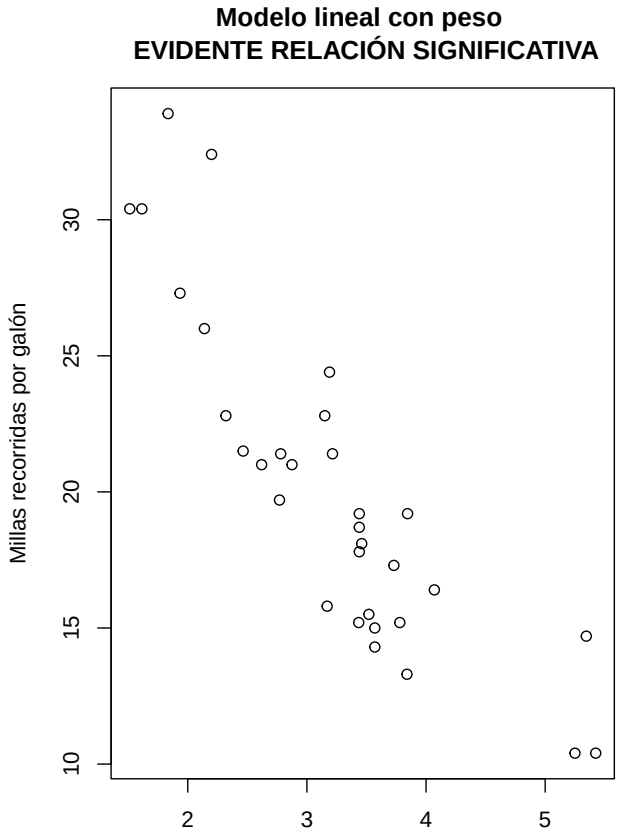
Note que incorporar una variable al modelo, esta puede o no ser relevante para explicar la variabilidad en Y. Si la variable Z no tiene efecto en la respuesta:

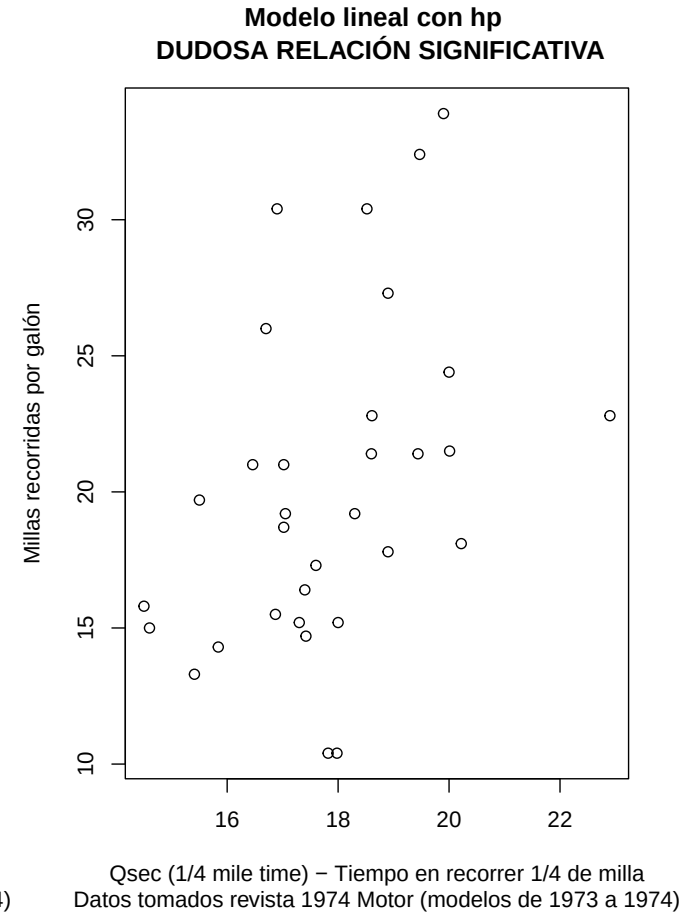
- Al minimizar SC_M, se lleva a $\beta_2 = 0$ y así $Y = \beta_0 + \beta_1 x + 0z = Y = \beta_0 + \beta_1 x$ tenemos el modelo de RLS.
- SC_M es la misma bajo los dos modelos, es decir añadir Z no mejoró, ni empeoró R^2.
- Añadir variables mantiene igual o incluso **mejora** R^2 aunque no sean de utilidad para explicar Y.
  - Variables irrelevantes pueden estar correlacionadas con la variable Y por coincidencia.
  - Mas variables irrelevantes aumentan la probabilidad de que esto suceda, mejorando artificialmente R^2.

En la práctica se reporta usualmente un R^2 ajustado por en número de variables.

Significancia de la regresión

Volviendo a RLS, no es claro cómo determinar o medir qué tan significativo resulta el valor de R^2, ¿en qué punto de R^2 el modelo es realmente mejor con o sin la covariable?





Recuerde que:

- La variación o error total en los datos, SC_T, corresponde a la variación total al asumir el modelo reducido.
- Al considerar la covariable mediante el modelo lineal disminuimos este error, la nueva variación la llamamos SC_E.
- La diferencia entre SC_T y SC_E corresponde a la variabilidad explicada por el modelo, o SC_M.

SC_T es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_T)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1,\ldots y_n$ necesarios para calcular SC_T:

- Para el cálculo de SC_T hacen faltan $gl(SC_T) = n - 1$ unidades para determinarlo:
  - $y'_1 = y_1 - \bar{y}$.
  - $\ldots$
  - $y'_{n-1} = y_n - \bar{y}$.
  - Como $\sum_i y'_i = 0$ se tiene $y'_n = -\sum_{i-1} y'_i$.
  - $y'_n = f(y'_1, ..., y'_{n-1})$.

Con lo cual $SC_T = \sum_i (y_i - \bar{y})^2 = \sum_i y_i'^2 = f(y'_1, ..., y'_{n-1})$ tiene $n - 1$ grados de libertad. Note que hace falta un parámetro (el promedio), para estimar SC_T, por eso se pierde un gl de los n que tienen los datos $y_1,\ldots,y_n$.

SC_E es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_E)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1, \ldots y_n$ necesarios para calcular SC_E:

- Para el cálculo de SC_E hacen falta $p = 2$ parámetros (pendiente e intercepto, en RLS) para ser estimado, por lo cual perdemos 2 grados de libertad, es decir $gl(SC_E) = n - p$.

SC_M es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_M)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1, \ldots y_n$ necesarios para calcular SC_M:

- Como vimos, $SC_M = SC_T - SC_E$, se tiene de la misma forma $gl(SC_M) = gl(SC_T) - gl(SC_E) = (n-1) - (n-p) = p - 1$, en RLS, $p - 1 = 2 - 1 = 1$.

De manera semejante a R^2, podemos definir una relación entre las sumas de cuadrados, esta vez entre SC_M y SC_E, para establecer un estadístico que caracterice la calidad del ajuste:

$$F' = \frac{SC_M}{SC_E}$$

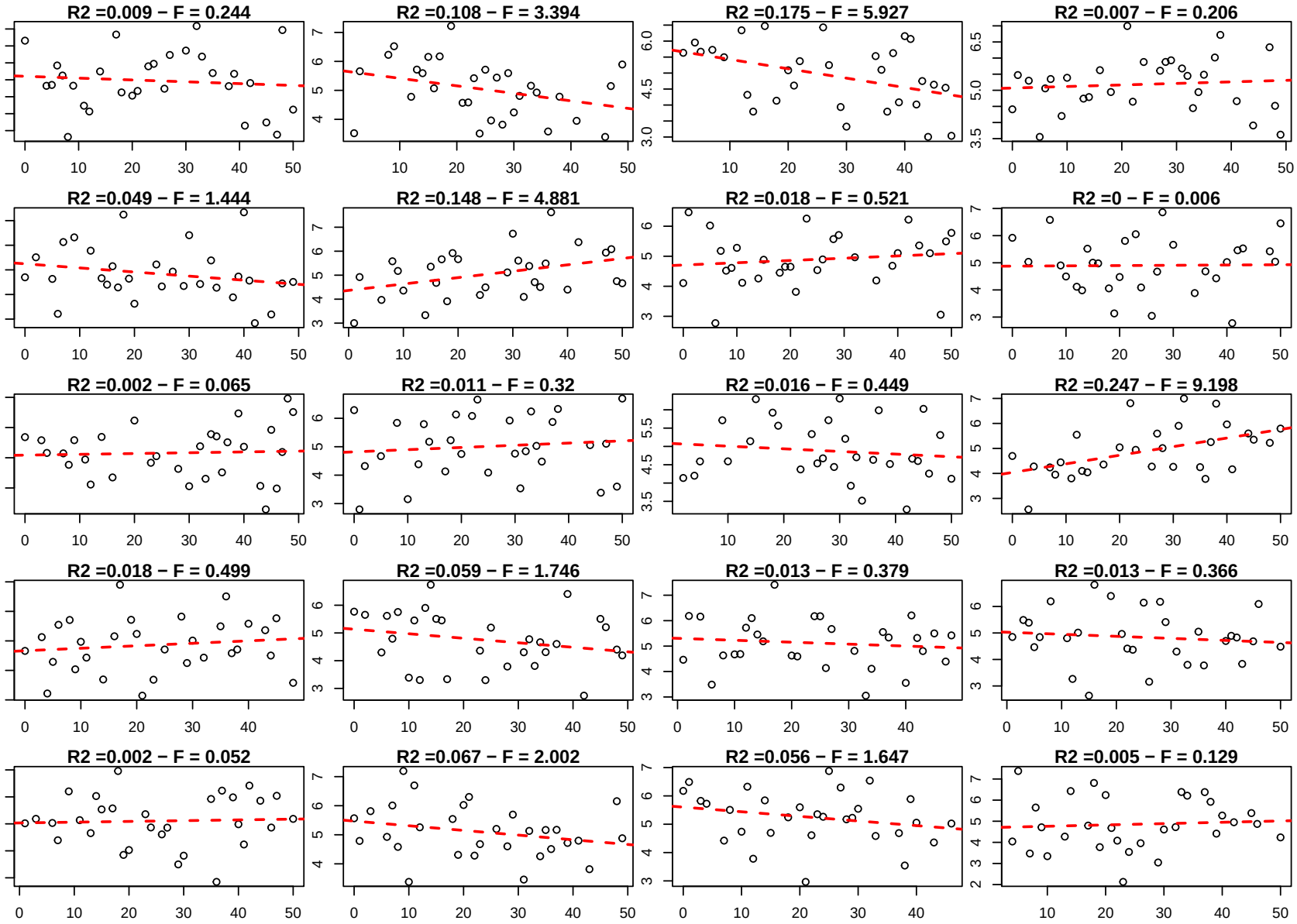
- F' es fácilmente interpretable: a medida que aumente, la variabilidad explicada por el modelo aumenta respecto a la que este deja de explicar.
- F' es una razón en lugar de una proporción, pero su diferencia con R^2 es de forma, más no de fondo. De hecho sus numeradores son iguales.

Al normalizar por los gl de cada SC en F', tenemos el estadístico F como un cociente de varianzas:

$$F = \frac{SC_M/gl(SC_M)}{SC_E/gl(SC_E)}$$

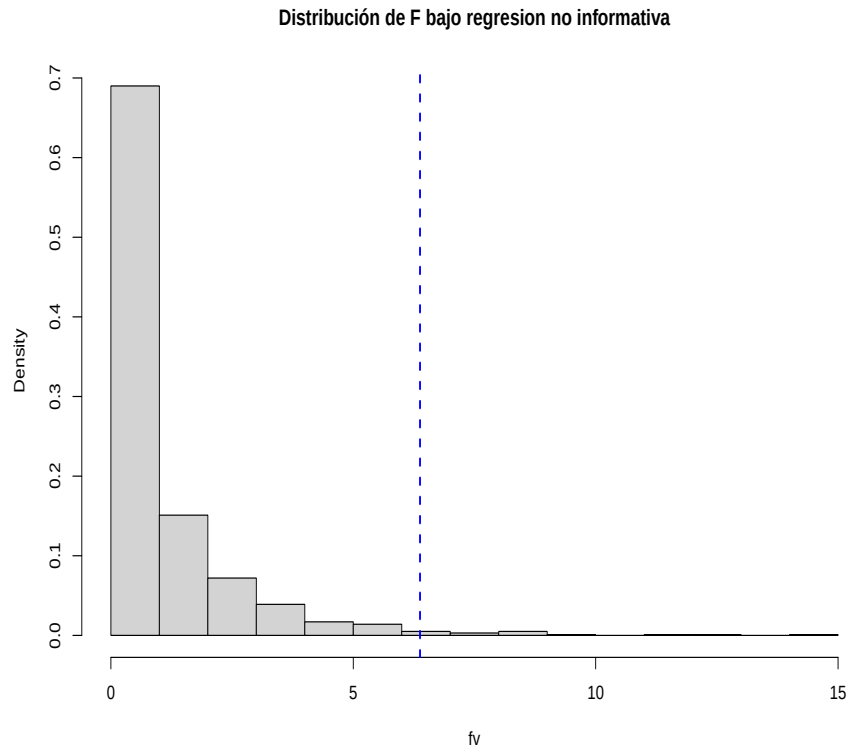
- Tanto F como F' tienen la misma interpretación.
- Su diferencia radica en que, cuándo la regresión no es informativa (es decir, los parámetros de las covariables son iguales a cero), F sigue una distribución estadística conocida, la distribución F-
- Los parámetros de F bajo la *hipótesis nula* son $gl(SC_M)$ en el numerador y $gl(SC_E)$ en el numerador y denominador.

Si la regresión no es informativa $\beta_1 = 0$ y los datos podrían verse como:



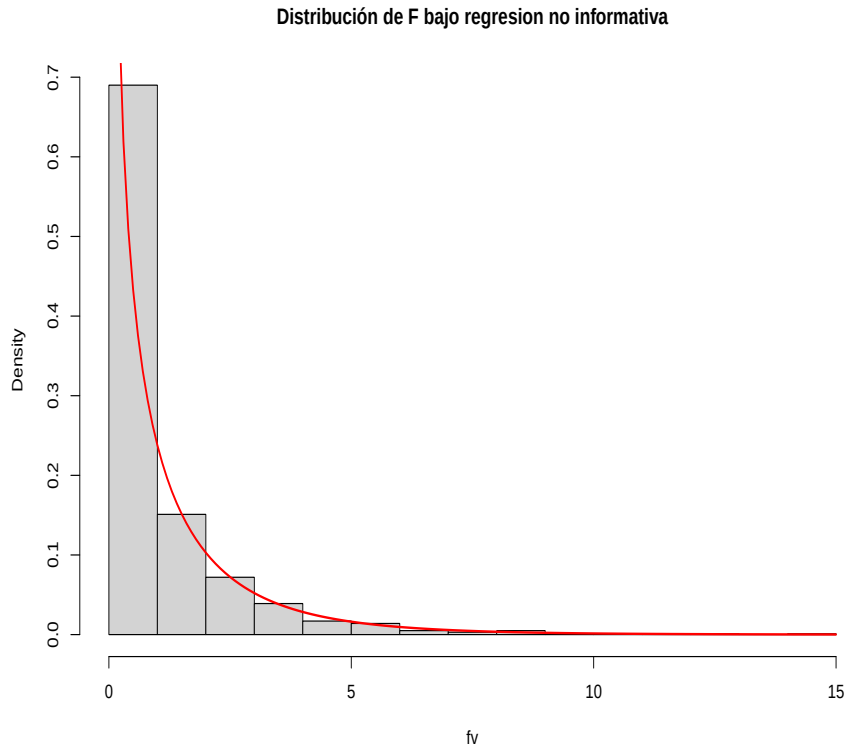
Podemos encontrar la distribución empírica de F bajo dicha **hipótesis** y agregar a esta los valores F observados para los modelos de regresión encontrados en la motivación de la sección

```
fv = NULL
set.seed(10)
for(i in 1:1000){
  xv    = sample(0:50,size=32) ; yv    = rnorm(32)
  lm_0  = lm(yv~xv)
  fv[i] = summary(lm_0)$fstatistic['value']
}
lm_sig  = lm(mpg~wt,data=mtcars)
lm_nsig = lm(mpg~qsec,data=mtcars)
hist(fv,prob=TRUE,main = 'Distribución de F bajo regresion no informativa')
abline(v = summary(lm_sig)$fstatistic['value'],col='red',lty=2,lwd=2)
abline(v = summary(lm_nsig)$fstatistic['value'],col='blue',lty=2,lwd=2)
```



No es necesario encontrar manualmente la distribución, ya que bajo la **hipótesis nula** F sigue una distribución estadística conocida, la distribución F, con $gl(SC_M) = 1$ en el numerador y $gl(SC_E) = n - 2 = 32 - 2 = 30$ en el denominador:

```
hist(fv,prob=TRUE,main = 'Distribución de F bajo regresion no informativa')
xval = seq(0,15,by=0.1)
yval = df(xval,df1=1,df2=30)
lines(xval,yval,col='red',lty=1,lwd=2)
```



Y con esta hacer inferencia (cálculo de p valor). Note que, nuevamente, esta cuantificación es también aplicable para RLM.

Se tiene que las hipótesis a probar son $H_0 : \beta_1 = ... = \beta_p = 0$ vs. H_1 : al menos un $\beta_j \neq 0$

```
anova(lm_sig)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73  91.375 1.294e-10 ***
## Residuals 30 278.32    9.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm_nsig)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value  Pr(>F)
## qsec       1 197.39 197.392  6.3767 0.01708 *
## Residuals 30 928.66  30.955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finalmente se destaca que hay una lista de premisas bajo el modelo lineal:

- Relación (aproximadamente) lineal.
- Error con media cero.
- Error con varianza constante.
- Errores no correlacionados - correlación bajo RLS/RLM implica una disminución artificial de la varianza - falsa significancia.
- Errores normalmente distribuídos - necesaria para probar, entre otras, la hipótesis sobre F.

Los cuales no son detectados mediante R^2, o F, al ser propiedades globales del modelo. Un modelo inadecuado puede resultar en conclusiones incluso opuestas a las reales bajo el proceso real generador de datos.