TEMA 3: Regresión

Parte 1: Regresión lineal

José R. Berrendero

Departamento de Matemáticas, Universidad Autónoma de Madrid

Temas a tratar

- El problema de regresión
- El error de predicción
- El modelo de regresión lineal
- Estimación e inferencia
- Comparación de modelos anidados
- Bootstrap en regresión

El problema de regresión

- Estudiar la relación entre una **variable respuesta** Y y un vector de p variables regresoras $X = (X_1, \ldots, X_p)$

- Suponemos varianzas finitas de todas las variables involucradas

- La relación entre $X_1, \ldots, X_p$ e Y se describe a través de la **función de regresión**

$$m(X) = \mathrm{E}(Y|X), \quad m(x) = \mathrm{E}(Y|X = x)$$

- Siempre se cumple que

$$Y = m(X) + \epsilon, \quad \mathrm{E}(\epsilon|X) = 0$$

  Basta definir $\epsilon = Y - m(X)$

- Recíprocamente, si $Y = m(X) + \epsilon$ con $\mathrm{E}(\epsilon|X) = 0$, entonces necesariamente $m(X) = \mathrm{E}(Y|X)$.

- A menudo se supone también la **hipótesis de homocedasticidad**:

$$\mathrm{Var}(\epsilon|X) = \mathrm{Var}(Y|X) = \sigma^2$$

El problema de regresión

- $m(X)$ es la mejor predicción de Y a partir de X, para cualquier valor de X: para cualquier función g tal que $g(X)$ tiene varianza finita

$$L_X(m) := \mathrm{E}[(Y - m(X))^2|X] \leq \mathrm{E}[(Y - g(X))^2|X] :=$$

- Pero $m(X)$ depende de la distribución de (Y, X), que es desconocida
- El objetivo general es estimar $m(\cdot)$ a partir de n observaciones iid de entrenamiento $(x_1, y_1), \ldots, (x_n, y_n)$, donde $x_i = (x_{i,1}, \ldots, x_{i,p})$.
- Los valores ajustados $\hat{Y}_1, \ldots, \hat{Y}_n$ se obtienen evaluando el estimador en $x_1, \ldots, x_n$, es decir, $\hat{Y}_i = \hat{m}(x_i)$
- Otro objetivo es predecir el valor de una nueva observación Y_0 desconocida a partir del correspondiente valor x_0. Se usa $\hat{Y}_0 = \hat{m}(x_0)$.

El problema de regresión

$$Y = m(X) + \epsilon$$

- $\mathrm{E}(\epsilon|X) = 0$ o, equivalentemente, $m(X) = \mathrm{E}(Y|X)$
- Homocedasticidad: $\mathrm{Var}(\epsilon|X) = \sigma^2$

Estas hipótesis implican $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$ y $\mathrm{Cov}(X, \epsilon) = 0$

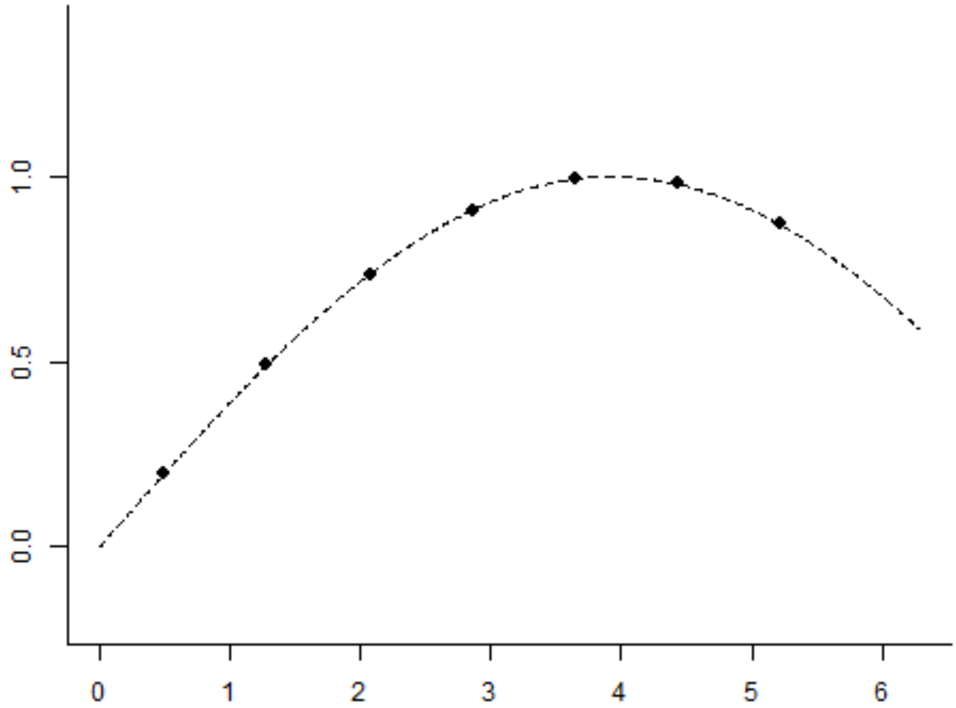
- **Modelos paramétricos:** m conocida salvo un número finito de parámetros. Ejemplo: regresión lineal

  $$m(X) = m(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

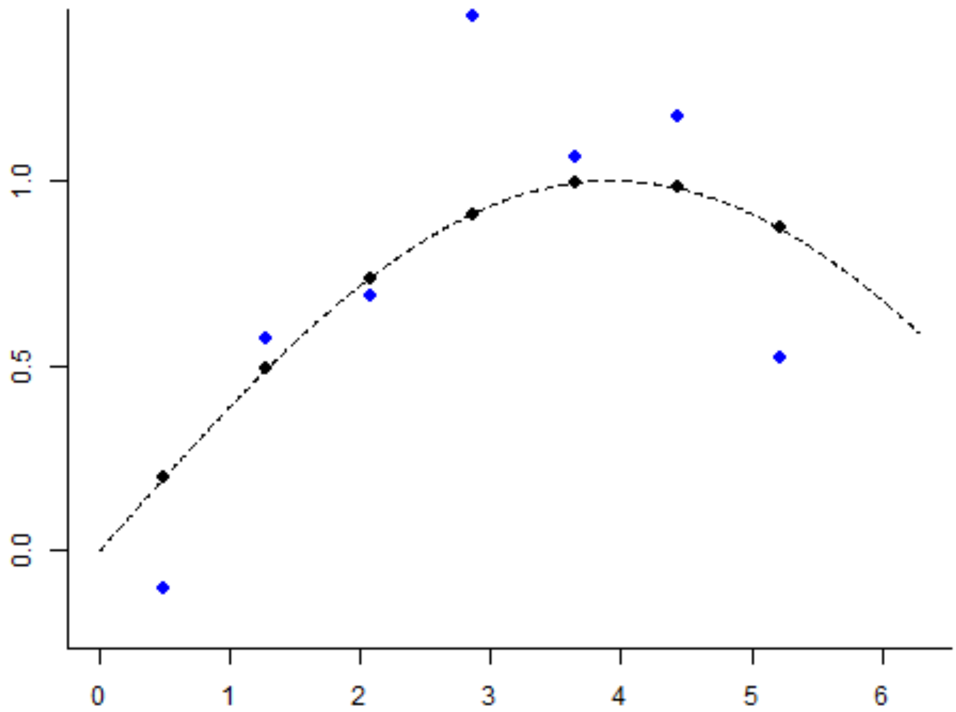
  - Si $p = 1$ *regresión lineal simple* y si $p > 1$ *regresión lineal múltiple*
  - Si Y también es un vector, *regresión lineal multivariante*

- **Modelos no paramétricos:** suelen suponer condiciones de continuidad o suavidad (existencia de derivadas) de la función $m(x)$.

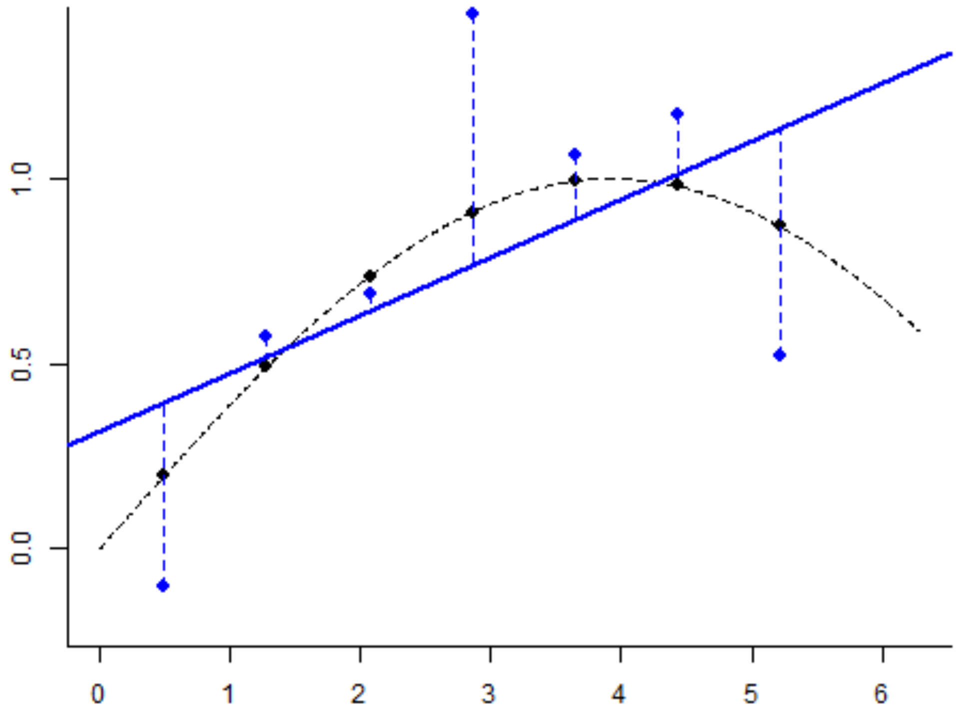
Modelo verdadero



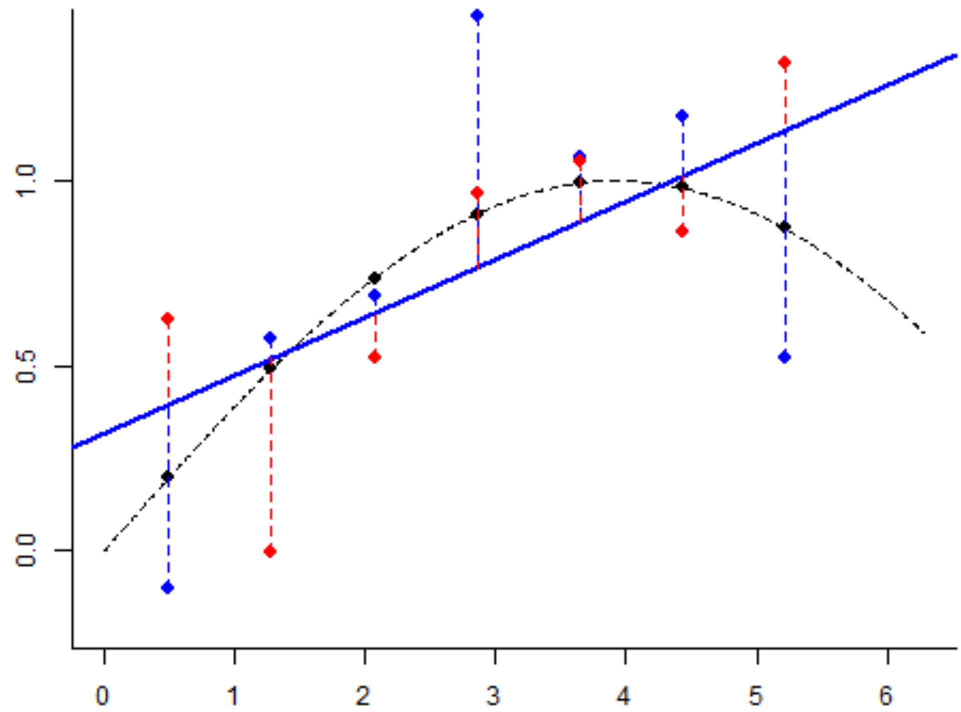
Datos de entrenamiento



Modelo ajustado (posiblemente falso)



Datos de test



Modelo verdadero

- Los valores $x_1, \ldots, x_n$ están prefijados y no se consideran aleatorios (diseño fijo)
- En el caso aleatorio, todas las medias y varianzas se deben entender condicionadas a las variables regresoras
- Muestra de entrenamiento (puntos azules) $Y = (Y_1, \ldots, Y_n)'$
- Muestra de test (puntos rojos) $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$

Hipótesis

- Vectores Y e $\tilde{Y}$ son i.i.d.
- Media: $\mathrm{E}(Y) = \mathrm{E}(\tilde{Y}) = \mu = (\mu_1, \ldots, \mu_n)'$ (puntos negros)

- No se supone nada sobre la verdadera relación entre X e Y, ni sobre el estimador utilizado

Error de entrenamiento y error de test

- Errores de entrenamiento y de test para la observación i:

$$\text{Training}(i) = \mathrm{E}[(Y_i - \hat{Y}_i)^2], \quad \text{Test}(i) = \mathrm{E}[(\tilde{Y}_i - \hat{Y}_i)^2$$

- Error de entrenamiento

$$\text{Training} = \frac{1}{n}\sum_{i=1}^{n} \text{Training}(i)$$

- Error de test

$$\text{Test} = \frac{1}{n}\sum_{i=1}^{n} \text{Test}(i)$$

- Optimismo: la diferencia entre el error de test y el error de entrenamiento

Error de entrenamiento y error de test

- Error de entrenamiento en i:

$$\mathrm{E}[(Y_i - \hat{Y}_i)^2] = \mathrm{E}[((Y_i - \mu_i) + (\mu_i - \hat{Y}_i))^2]$$
$$\mathrm{E}[(Y_i - \hat{Y}_i)^2] = \mathrm{Var}(Y_i) + \mathrm{E}[(\mu_i - \hat{Y}_i)^2] + 2\mathrm{E}[(Y_i - \mu_i)(\mu_i$$

- Error de test en i:

$$\mathrm{E}[(\tilde{Y}_i - \hat{Y}_i)^2] = \mathrm{E}[((\tilde{Y}_i - \mu_i) + (\mu_i - \hat{Y}_i))^2] = \mathrm{Var}(Y_i)$$

- Por lo tanto, el optimismo para la observación i es

$$\mathrm{Test}(i) - \mathrm{Training}(i) = 2\mathrm{E}[(Y_i - \mu_i)(\hat{Y}_i - \mu_i)] = 2\mathrm{Co}$$

ya que

$$\mathrm{E}[(Y_i - \mu_i)(\hat{Y}_i - \mu_i)] = \mathrm{E}[(Y_i - \mu_i)\hat{Y}_i] = \mathrm{Cov}(Y_i, \hat{Y}_i)$$

Error de entrenamiento y error de test

$$\text{Test} = \text{Training} + \frac{2}{n}\sum_{i=1}^{n} \text{Cov}(Y_i, \hat{Y}_i)$$

- Una covarianza alta entre las respuestas y los valores ajustados es un indicador de sobreajuste, lo que lleva a que el errorde entrenamiento pueda infraestimar seriamente el error de predicción
- **Ejercicio:** supongamos que a falta de más información usamos la media para predecir, esto es, $\hat{Y}_i = \bar{Y}$, para todo $i = 1, \ldots, n$. ¿Cuál es el valor del optimismo en este caso?

El modelo de regresión lineal múltiple

El modelo de regresión más usual es el siguiente:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

donde $\mathrm{E}(\epsilon|X_1, \ldots, X_p) = 0$ y $\mathrm{Var}(\epsilon|X_1, \ldots, X_p) = \sigma^2$.

Para muchas inferencias (intervalos, contrastes, etc.) se asume además que $\epsilon|(X_1, \ldots, X_p)$ tiene distribución normal.

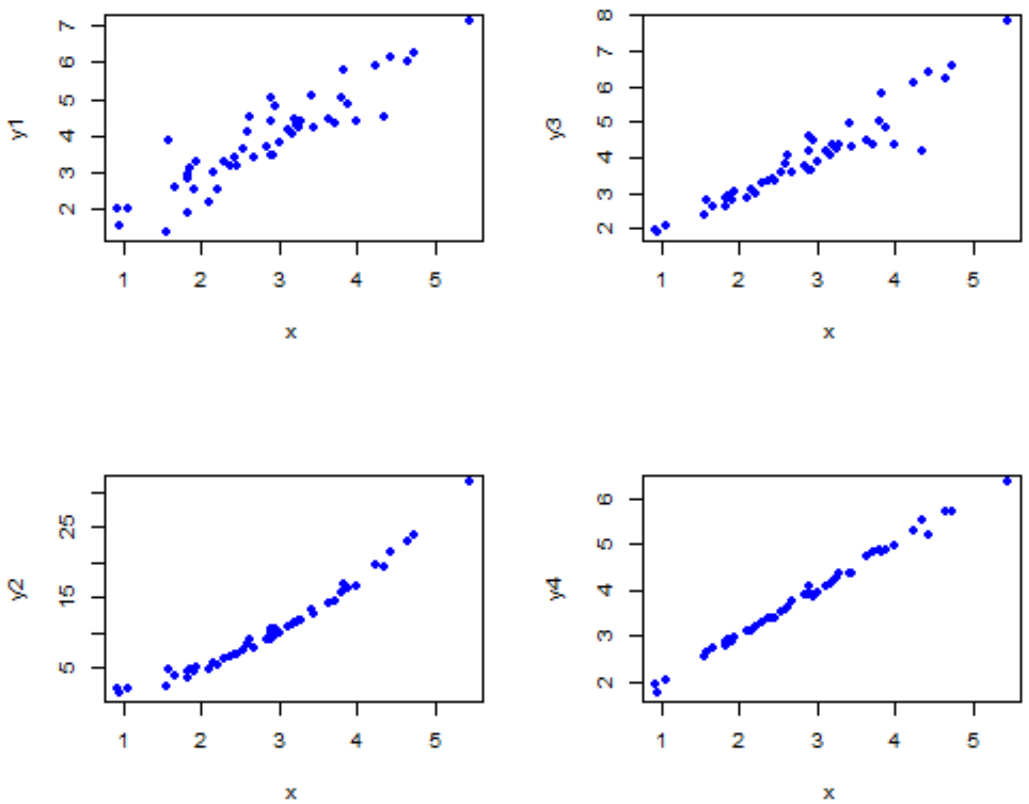
Cada observación de la muestra de entrenamiento sigue el modelo

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i, \quad i = 1, \ldots, n.$$

donde $\mathrm{E}(\epsilon_i|x_i) = 0$ y $\mathrm{Var}(\epsilon_i|x_i) = \sigma^2$.

El modelo de regresión lineal múltiple

¿Cuáles de los siguientes conjuntos de datos verifican el modelo?



El modelo de regresión lineal múltiple

En forma matricial,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ 1 & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

De forma más compacta,

$$Y = X\beta + \epsilon, \quad \epsilon|X \equiv \mathrm{N}_n(0, \sigma^2 \mathbb{I}_n) \Leftrightarrow Y|X \equiv \mathrm{N}_n(X\beta, \sigma^2 \mathbb{I}_n)$$

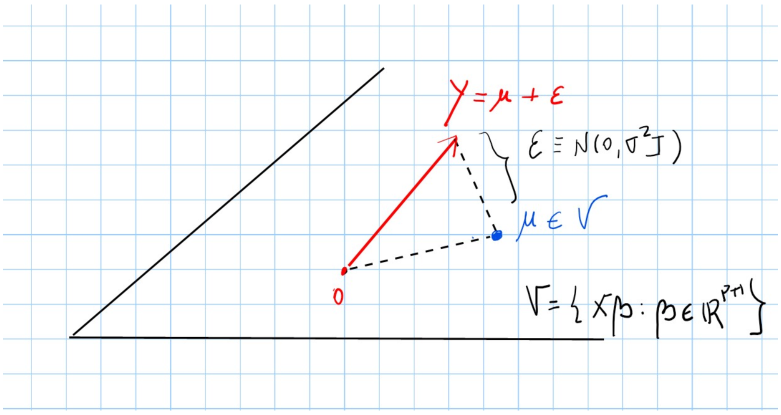
X es la **matriz de diseño**

Interpretación geométrica

Llamaremos $V = R(X) \subset \mathbb{R}^n$ al subespacio vectorial generado por las columnas de la matriz de diseño X

$$\mu \in V \Leftrightarrow \text{Existe } \beta \in \mathbb{R}^{p+1} \text{ tal que } \mu = X\beta$$

El modelo de regresión equivale a $Y|X \equiv \mathrm{N}_n(\mu, \sigma^2 \mathbb{I}_n)$, donde $\mu \in V$.



Ajuste por mínimos cuadrados

- Estimadores de mínimos cuadrados: los coeficientes $\hat{\beta}_0, \ldots, \hat{\beta}_p$ para los que se minimiza

$$\|Y - X\beta\|_2^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p})]^2$$

- $\hat{Y} \equiv X\hat{\beta}$ es la **proyección ortogonal** de Y sobre V.

- Ecuaciones normales: el vector $e = Y - \hat{Y} = Y - X\hat{\beta}$ se denomina **vector de residuos**. Los residuos deben ser ortogonales a las columnas de X (una base de V):

$$X'(Y - \hat{Y}) = 0 \Leftrightarrow X'e = 0$$

- Expresión de los estimadores de mínimos cuadrados:

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

Ajuste por mínimos cuadrados

- El estimador de mínimos cuadrados es el estimador de máxima verosimilitud (EMV) de β:

$$L(\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right\}$$

- El vector $\hat{\beta}$ tiene distribución normal $(p+1)$-dimensional con vector de medias β y matriz de covarianzas $\sigma^2(X'X)^{-1}$:

$$\hat{\beta} \equiv \mathrm{N}_{p+1}(\beta, \sigma^2(X'X)^{-1})$$

- El vector de **valores ajustados** es

$$\hat{Y} = X\hat{\beta} = HY, \quad H = X(X'X)^{-1}X'$$

  A H se le llama *matriz sombrero*, y geométricamente es una matriz de proyección sobre V

- El vector de residuos es entonces $e = Y - \hat{Y} = (I - H)Y$

Ajuste por mínimos cuadrados

- Para estimar la varianza σ^2 se usa la **varianza residual**

$$S_R^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

  ¿Por qué estos son los grados de libertad apropiados?

- Un resultado importante es que $(n-p-1)S_R^2/\sigma^2 \equiv \chi^2_{n-p-1}$, lo que permite construir intervalos de confianza y contrastes para σ^2

- Además, S_R^2 y $\hat{\beta}$ son independientes

Los resultados que acabamos de resumir son la base para obtener los intervalos y contrastes para los coeficientes del modelo

Descomposición de la variabilidad

- **Suma de cuadrados total:** $\text{SCT} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, mide la variabilidad total en la respuesta.
- **Suma de cuadrados explicada:** $\text{SCE} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, mide la parte de la variabilidad explicada por el modelo.
- **Suma de cuadrados residual:** $\text{SCR} = \sum_{i=1}^{n} e_i^2$, mide la parte de la variabilidad no explicada por el modelo.

Usando la ortogonalidad de los residuos con las variables regresoras

$$\text{SCT} = \text{SCE} + \text{SCR}$$

La variabilidad total de la respuesta se puede descomponer en una parte explicada por las variables regresoras y otra no explicada o residual

Descomposición de la variabilidad

- El **coeficiente de determinación** es una medida de la capacidad del modelo para explicar Y:

$$R^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}}$$

- **Contraste de la regresión:** Para contrastar $H_0 : \beta_1 = \cdots = \beta_p = 0$ se usa

$$F = \frac{\mathrm{SCE}/p}{\mathrm{SCR}/(n-p-1)}$$

Bajo H_0, el estadístico F sigue una distribución $F_{p,n-p-1}$

Ajuste del modelo con R

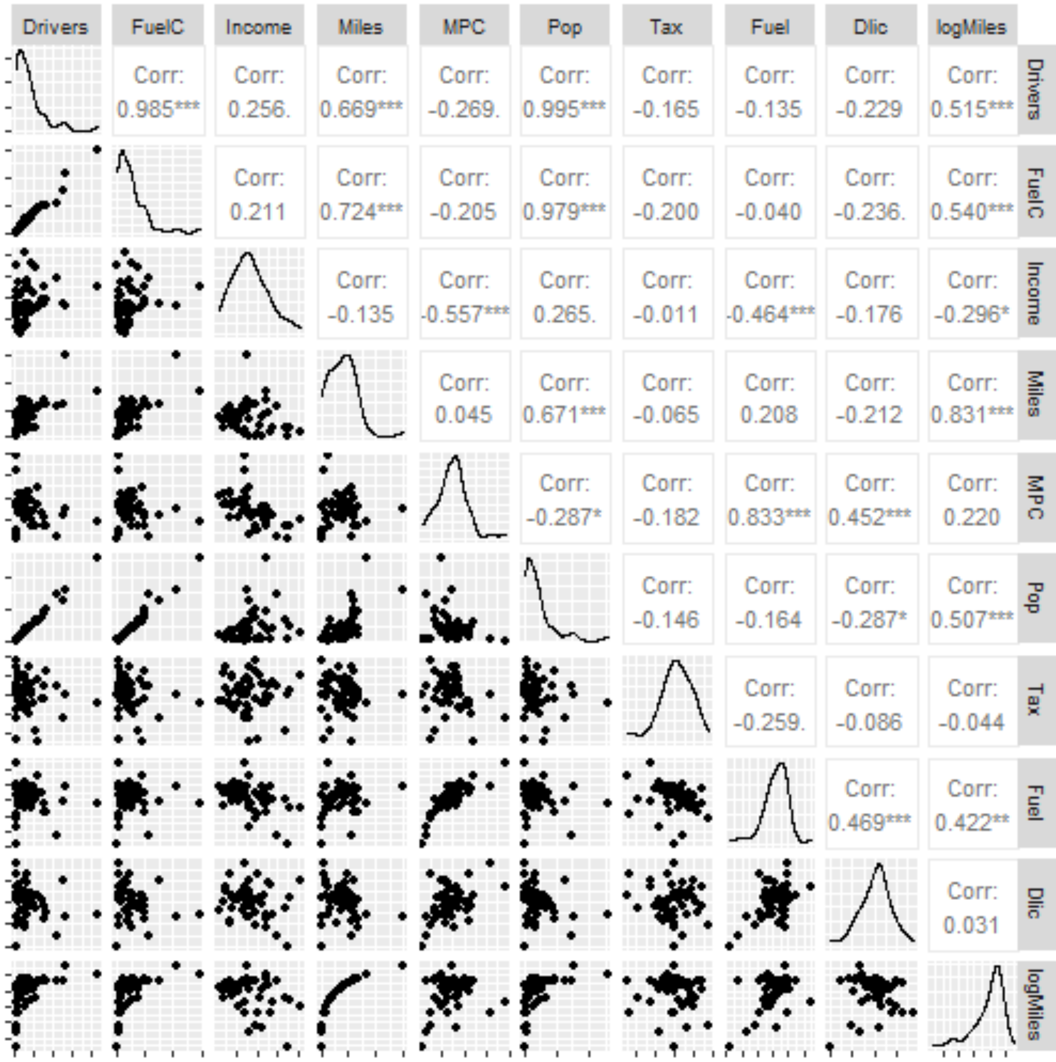
- Datos `fuel2001`: consumo de combustible (y otras variables relacionadas) en EE.UU.

```
datos <- 'http://verso.mat.uam.es/~joser.berrendero/datos
load(url(datos))
head(fuel2001)
```

```
##      Drivers    FuelC Income  Miles      MPC      Pop  Tax
## AL   3559897  2382507  23471  94440 12737.00  3451586 18.0
## AK    472211   235400  30064  13628  7639.16   457728  8.0
## AZ   3550367  2428430  25578  55245  9411.55  3907526 18.0
## AR   1961883  1358174  22257  98132 11268.40  2072622 21.7
## CA  21623793 14691753  32275 168771  8923.89 25599275 18.0
## CO   3287922  2048664  32949  85854  9722.73  3322455 22.0
```

```
fuel2001 <- fuel2001 %>%
  mutate(Fuel = 1000 * FuelC/Pop) %>%
  mutate(Dlic = 1000 * Drivers/Pop) %>%
  mutate(logMiles = log(Miles))

# Diagramas de dispersión
library(GGally)
ggpairs(fuel2001) +
  theme(axis.text=element_blank())
```

Ajuste del modelo con R



Ajuste del modelo con R

```
reg <- lm(Fuel ~ Tax + Dlic + Income + logMiles,
          data=fuel2001)
summary(reg)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.192845 194.906161   0.791 0.432938
## Tax          -4.227983   2.030121  -2.083 0.042873 *
## Dlic          0.471871   0.128513   3.672 0.000626 ***
## Income       -0.006135   0.002194  -2.797 0.007508 **
## logMiles     26.755176   9.337374   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105,    Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Ajuste del modelo con R

```
nuevo.dato <- data.frame(18, 1031, 23471, 11)
names(nuevo.dato) <- names(fuel2001)[c(7, 9, 3, 10)]
nuevo.dato
```

```
##   Tax Dlic Income logMiles
## 1  18 1031  23471       11
```

```
predict(reg, nuevo.dato, interval='confidence')
```

```
##        fit      lwr      upr
## 1 714.8929 674.1173 755.6686
```

```
predict(reg, nuevo.dato, interval='prediction')
```

```
##        fit      lwr      upr
## 1 714.8929 578.0571 851.7288
```

Error de predicción

- Recordamos la fórmula general $\text{Test} = \text{Training} + 2/n \sum_{i=1}^{n} \text{Cov}(Y_i, \hat{Y}_i)$
- Sea $\text{Cov}(Y, \hat{Y})$ la matriz $n \times n$ cuyas entradas son $\text{Cov}(Y_i, \hat{Y}_j)$
- En regresión lineal, $\hat{Y} = HY$, por lo tanto

$$\sum_{i=1}^{n} \text{Cov}(Y_i, \hat{Y}_i) = \text{traza}[\text{Cov}(Y, \hat{Y})] = \text{traza}[\text{Cov}(Y, H$$

donde Σ es la matriz de covarianzas de Y (condicionada a X si los regresores son aleatorios)

$$\text{Test} = \text{Training} + \frac{2}{n} \sum_{i=1}^{n} \text{traza}(\Sigma H)$$

Error de predicción

Variables independientes y homocedásticas

- Las hipótesis más habituales:

$$\Sigma = \sigma^2 \mathbb{I}_n \Rightarrow \text{traza}(\Sigma H) = \sigma^2 \text{traza}(H) = \sigma^2 (p+1)$$

$$\text{Test} = \text{Training} + \frac{2}{n}(p+1)\sigma^2$$

- El valor $p+1$ corresponde al caso en que hay término independiente y p variables regresoras. En general la traza es la dimensión del espacio sobre el que se proyecta.

- Habíamos supuesto diseño fijo (o resultado condicionado a los valores de X) pero el resultado no depende de estos valores, por lo tanto es también válido para regresores aleatorios

Error de predicción

Variables independientes pero heterocedásticas

- $\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$

$$\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\} \Rightarrow \text{traza}(\Sigma H) = \sum_{i=1}^n \sigma_i^2 h_{ii},$$

donde h_{ii} es el *leverage* de la observación i, el correspondiente elemento de la diagonal de H

$$\text{Test} = \text{Training} + \frac{2}{n} \sum_{i=1}^n \sigma_i^2 h_{ii}$$

Cuestión: ¿Cómo se extiende este resultado al caso de regresores aleatorios?

Comparación de modelos anidados

- Un modelo complejo se ajusta mejor a los datos disponibles pero ello no significa que proporcione mejores predicciones (ya que el optimismo crece con p)
- Un modelo sencillo evita el sobreajuste pero puede introducir sesgos
- Objetivo: comparar dos modelos lineales tales que uno es una simplificación del otro contrastando $H_0 : A\beta = 0$, donde A es una matriz $k \times (p+1)$ con rango$(A) = k < p+1$.
- Por ejemplo, en el modelo $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, queremos contrastar

  $$H_0 : \beta_1 = \beta_2; \beta_0 = 0 \Leftrightarrow A\beta = 0$$

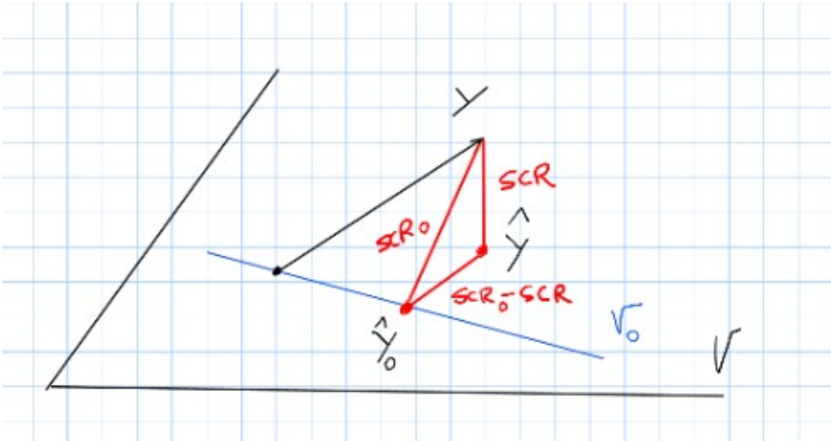
  donde

  $$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix}$$

El **modelo reducido** (M0) es el que resulta de imponer las restricciones de H_0

Interpretación geométrica

El modelo reducido equivale a $\mu \in V_0 \subset V$, donde V_0 es un supespacio de V de dimensión $p + 1 - k$



La idea básica

- SCR_0 es la variabilidad no explicada (residual) bajo el modelo reducido
- SCR es la variabilidad no explicada (residual) bajo el modelo completo
- Siempre se cumple $\mathrm{SCR}_0 > \mathrm{SCR}$ (¿por qué?) Se rechaza H_0 si

$$\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}}$$

  es suficientemente grande (si complicar el modelo merece la pena)

- Bajo $H_0 : A\beta = 0$, se verifica

$$\frac{(\mathrm{SCR}_0 - \mathrm{SCR})/k}{\mathrm{SCR}/(n-p-1)} \equiv F_{k,n-p-1}$$

  La región crítica del contraste para un nivel α es

$$R = \left\{ \frac{(\mathrm{SCR}_0 - \mathrm{SCR})/k}{\mathrm{SCR}/(n-p-1)} > F_{k,n-p-1;\alpha} \right\}$$

Comparación con R

Se ajustan ambos modelos:

```r
# Modelo completo
reg <- lm(Fuel ~ Tax + Dlic + Income + logMiles,
          data=fuel2001)

# Modelo reducido
reg0 <- lm(Fuel ~ logMiles, data=fuel2001)
anova(reg0)
```

```
## Analysis of Variance Table
##
## Response: Fuel
##           Df Sum Sq Mean Sq F value   Pr(>F)
## logMiles   1  70478   70478  10.619 0.002038 **
## Residuals 49 325216    6637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparación con R

Se comparan las sumas de cuadrados residuales usando el comando `anova`

```
anova(reg0, reg)
```

```
## Analysis of Variance Table
##
## Model 1: Fuel ~ logMiles
## Model 2: Fuel ~ Tax + Dlic + Income + logMiles
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     49 325216
## 2     46 193700  3    131516 10.411 2.402e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{k}}{\frac{\mathrm{SCR}}{n-p-1}} = \frac{\frac{325216-193700}{3}}{\frac{193700}{46}} = 10.411$$

El contraste $H_0 : \beta_1 = \cdots = \beta_p = 0$ es un caso particular de comparación entre un modelo reducido y un modelo completo

Bootstrap en regresión

- Regresión simple, diseño fijo
- La varianza de la pendiente de la recta de mínimos cuadrados es

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Supongamos que no conocemos o no sabemos deducir la fórmula anterior. El objetivo es aproximar $\mathrm{Var}(\hat{\beta}_1)$ mediante simulación.
- El método bootstrap también da una aproximación a la distribución de $\hat{\beta}_1$ en el caso en que no se cumple la hipótesis de normalidad

Ejemplo: generación de datos

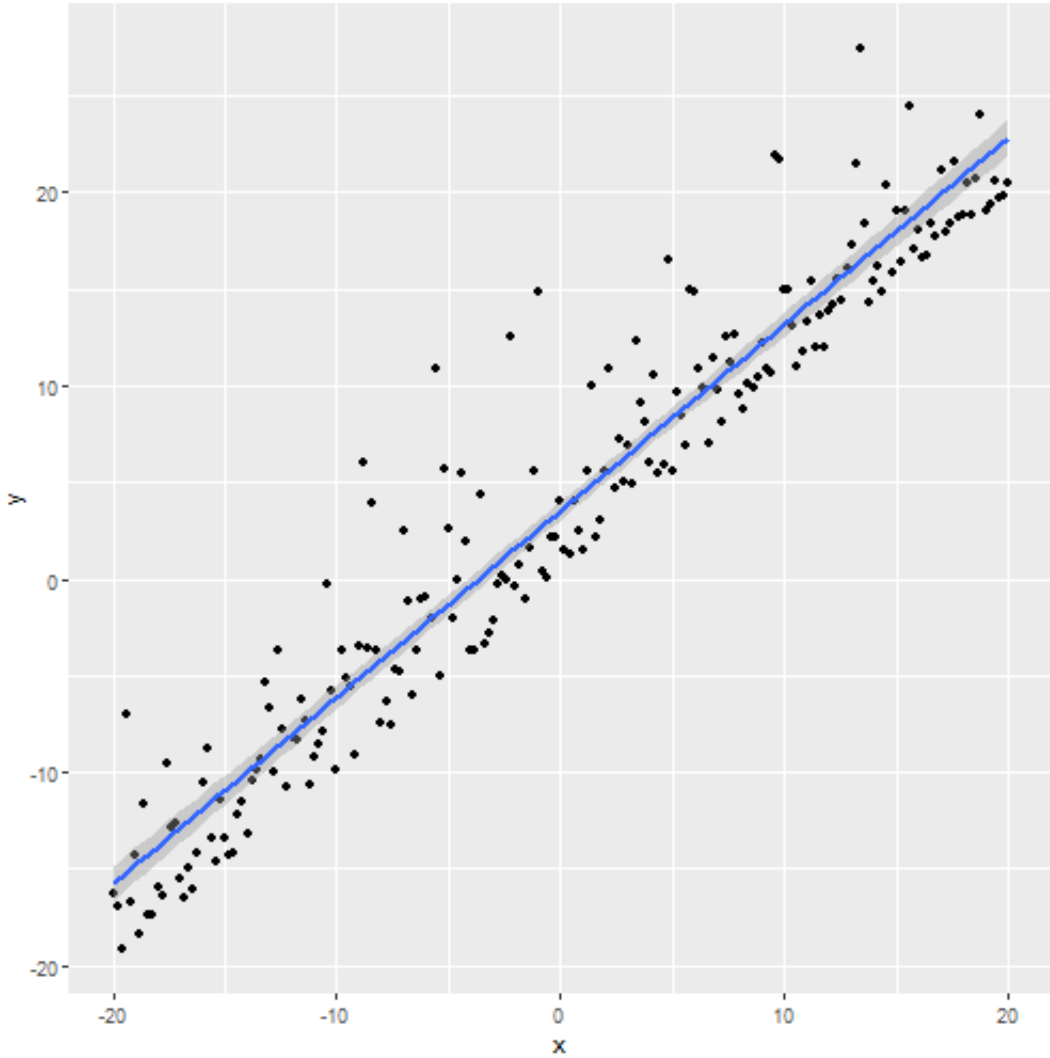
```
# Parámetros
beta0 <- 0
beta1 <- 1
sigma <- 4

# Muestra original de (x, y)
set.seed(100)
x <- seq(-20, 20, 0.2)
n <- length(x)

epsilon <- rexp(n, rate = 1/sigma)
y <- beta0 + beta1*x + epsilon
summary(lm(y~x))

# Desviación típica verdadera
dt_beta1 <- sigma / sqrt(sum((x-mean(x))^2))
dt_beta1

# Representación gráfica
ggplot(data.frame(x, y), aes(x, y)) +
  geom_point() +
  geom_smooth(method = lm)
```

Ejemplo: generación de datos

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.846 -2.329 -1.083  1.192 12.780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.53993    0.23729   14.92   <2e-16 ***
## x            0.96474    0.02045   47.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.364 on 199 degrees of freedom
## Multiple R-squared:  0.9179,    Adjusted R-squared:  0.9175
## F-statistic:  2226 on 1 and 199 DF,  p-value: < 2.2e-16
```

```
## [1] 0.02431263
```

Ejemplo: generación de datos



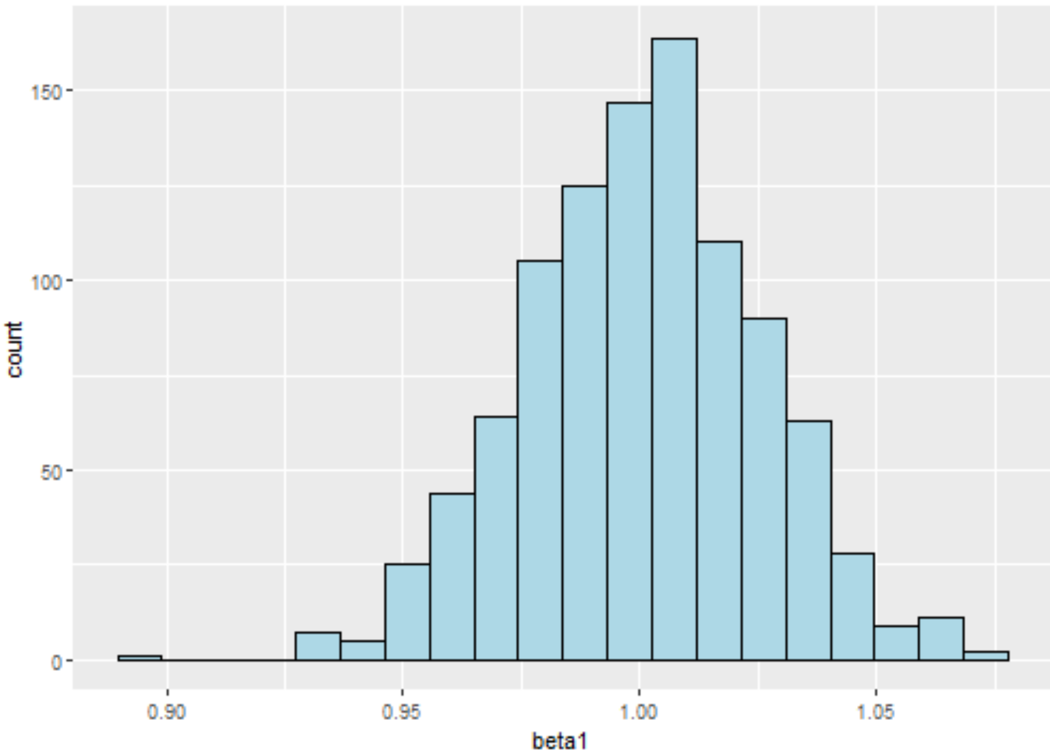
Ejemplo: simulación

```
# Aproximación por simulación
R <- 1000

beta1_sim <- NULL
for (i in 1:R){
  epsilon_sim <- rexp(n, rate = 1/sigma)
  y_sim <- beta0 + beta1*x + epsilon_sim
  beta1_sim[i] <- coefficients(lm(y_sim ~ x))[2]
}
sd(beta1_sim)
```

```
## [1] 0.02497163
```

Ejemplo: simulación

```
ggplot(data.frame(beta1 = beta1_sim)) +
  geom_histogram(aes(x = beta1), fill = 'lightblue', col
```



Ejemplo: bootstrap

- Como no conocemos los parámetros usamos los valores estimados con la muestra original
- Como no sabemos cuál es la distribución de los errores, generamos los valores de Y usando la distribución empírica de los residuos

```
# Aproximación por bootstrap
R <- 1000

# Parámetros estimados
reg <- lm(y~x)
beta0_hat <- coefficients(reg)[1]
beta1_hat <- coefficients(reg)[2]
residuos <- residuals(reg)


beta1_boot <- NULL
for (i in 1:R){
  epsilon_boot <- sample(residuos, n, rep = TRUE)
  y_boot <- beta0_hat + beta1_hat*x + epsilon_boot
  beta1_boot[i] <- coefficients(lm(y_boot ~ x))[2]
}
sd(beta1_boot)
```

```
## [1] 0.01964217
```

Ejemplo: bootstrap

```
ggplot(data.frame(beta1 = beta1_boot)) +
  geom_histogram(aes(x = beta1), fill = 'lightblue', col
```