Elementos de estadística para análisis de expresión diferencial

Sesión 1.4

Alejandro Reyes

Octubre 24, 2016

Temas

- Modelos lineales
- P-value
- Problema de multiplicidad de pruebas
- Fuentes de variación en experimentos de RNA-seq

¿Qué es un regresión lineal?

- Una regresión lineal modela la respuesta de una variable Y como una combinación lineal de variables predictoras.

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

- Y: es la variable a predecir o la variable dependiente
- X: es un vector de predictores o variables independientes
- β_p: son los coeficientes de la regresión, los parámetros a estimar
- ε: es el error del modelo

Ejemplo de una regresión lineal I

Por vecindario

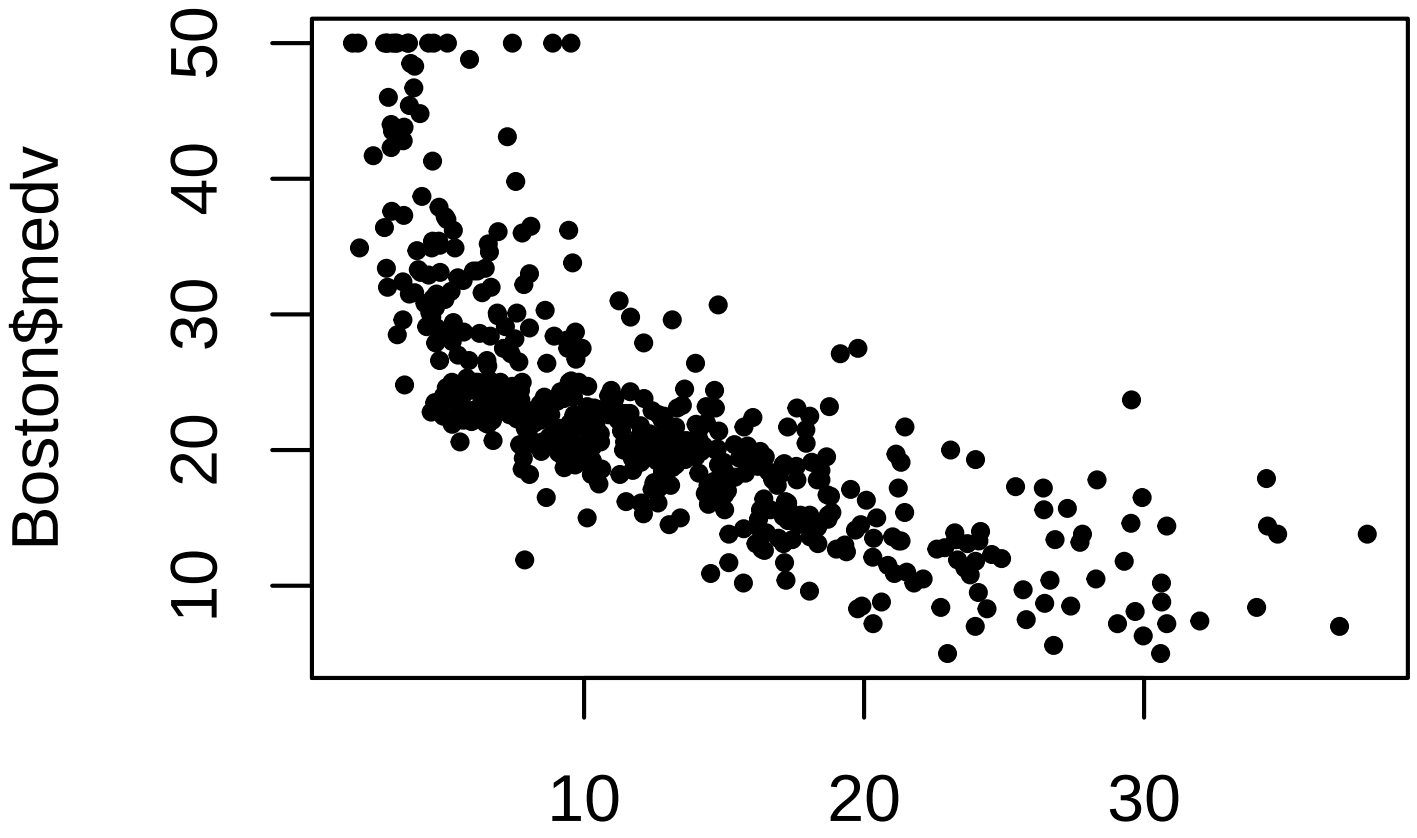
- **lstat** es el porcentage de casas con nivel socio económico bajo
- **medv** es el valor medio de una casa
- **age** es la edad media de las casas

```
library(MASS)
data("Boston")
head( Boston[,c("lstat", "medv", "age")])
```

```
##   lstat medv  age
## 1  4.98 24.0 65.2
## 2  9.14 21.6 78.9
## 3  4.03 34.7 61.1
## 4  2.94 33.4 45.8
## 5  5.33 36.2 54.2
## 6  5.21 28.7 58.7
```

Ejemplo de una regresión lineal II

```
plot(Boston$lstat, Boston$medv, pch=19, cex=.5)
```



Ejemplo de una regresión lineal III

$medv \approx \beta_0 + \beta_1 lstat$

```
modelo <- lm( medv ~ lstat, Boston )
modelo
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Coefficients:
## (Intercept)        lstat
##       34.55        -0.95
```

- β_0: media de **medv**
- β_1: como cambia **medv** cada vez que incrementamos por una unidad **lstat**

Ejemplo de una regresión lineal IV

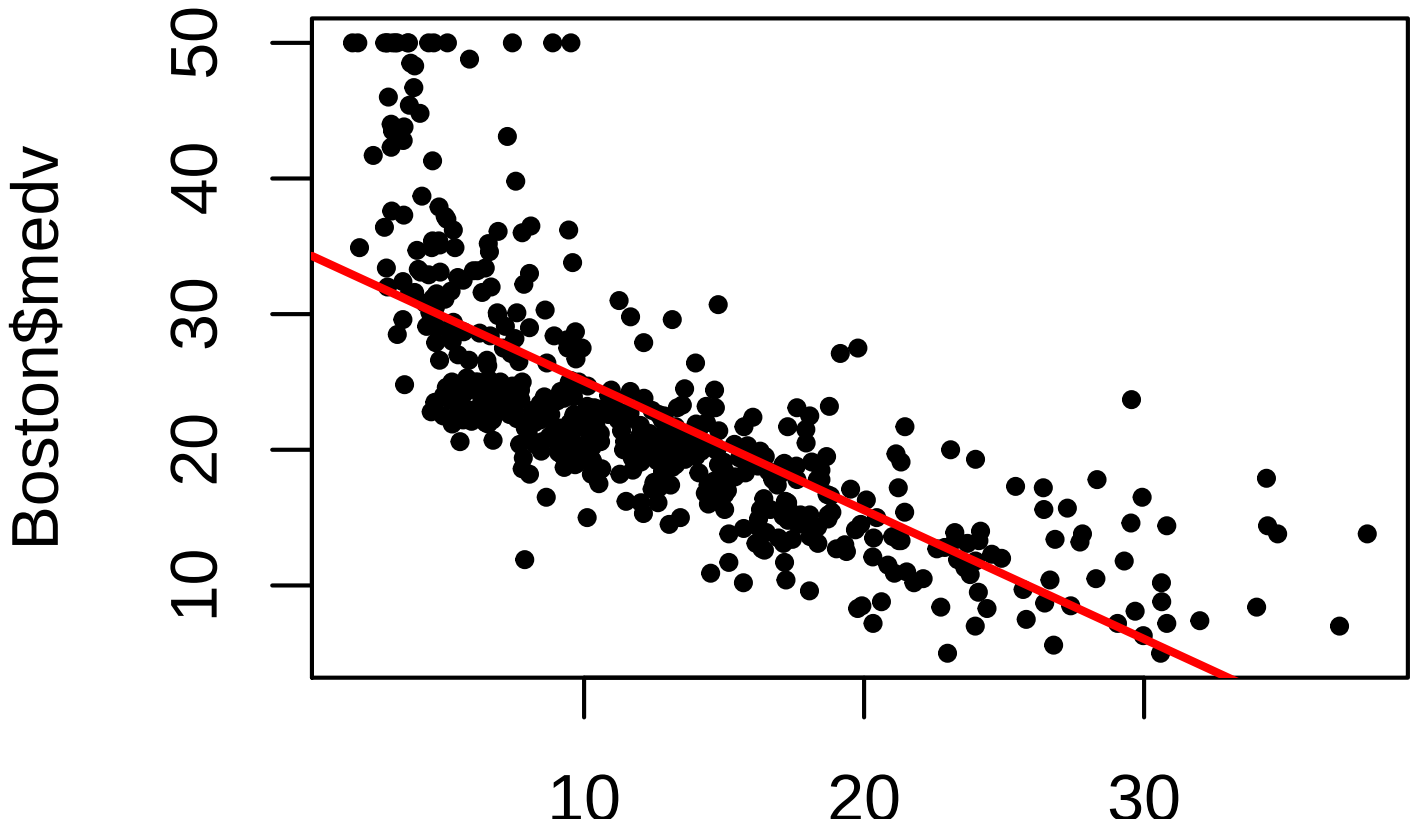
¿Cuál es el precio medio que el modelo predice si el % de casas con nivel socioeconómico bajo es del 15%?

```
coefs <- coefficients( modelo )
coefs["(Intercept)"] + coefs["lstat"] * 15
```

```
## (Intercept)
##     20.3031
```

Ejemplo de una regresión lineal V

```
plot(Boston$lstat, Boston$medv, pch=19, cex=.5)
abline(modelo, lwd=2, col="red")
```



Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$medv \approx \beta_0 + \beta_1 lstat + \beta_2 age$$

```
modelo2 <- lm( medv ~ lstat + age, Boston )
modelo2
```

```
##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Coefficients:
## (Intercept)        lstat          age
##    33.22276     -1.03207      0.03454
```

Ejercicio I

¿Cuál es el precio medio estimado por el modelo si el porcentage de casas con nivel socioeconómico bajo es del 10% y si la edad media de las casas es de 20 años?

Regresiones con variables cualitativas I

```
library(ISLR)
data( Wage )
head( Wage[,c("wage", "sex", "education", "health")] )
```

```
##                wage     sex       education         health
## 231655   75.04315 1. Male    1. < HS Grad      1. <=Good
## 86582    70.47602 1. Male 4. College Grad 2. >=Very Good
## 161300 130.98218 1. Male 3. Some College      1. <=Good
## 155159 154.68529 1. Male 4. College Grad 2. >=Very Good
## 11443    75.04315 1. Male      2. HS Grad      1. <=Good
## 376662 127.11574 1. Male 4. College Grad 2. >=Very Good
```

Regresiones con variables cualitativas II

Uso de "dummy variables" para codificar los factores cualitativos

$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon$

$x_i = \{$ 0 si la persona i es hombre, 1 si la persona i es mujer $\}$

$y_i = \{$ β_0 si x_i es hombre,

$\beta_0 + \beta_1$ si x_i es mujer $\}$

Variación en cuentas de *RNA-seq*

* Variación = variación técnica + variación biológica

Variación técnica en cuentas de *RNA-seq* (Poisson) I

Asuman que tienen un costal con pelotas, 90% son blancas y el 10% son rojas. A 50 personas (que no saben lo anterior) les dan la tarea de estimar el porcentage de pelotas blancas haciendo una muestra de n pelotas.

- ¿Cuál es el valor estimado de pelotas rojas si $n = 20$?

```
obtenerValorEsperado <- function( fracPelotasBlancas, n )
{
    valorEsperado <- n * fracPelotasBlancas
    valorEsperado
}
n <- 20
fracPelotasBlancas <- .1
valorEsperado <-
  obtenerValorEsperado( fracPelotasBlancas, n )
valorEsperado
```

```
## [1] 2
```

Variación técnica en cuentas de *RNA-seq* (Poisson) II

Los 5 de ustedes hacen un sampleo y obtienen los siguentes números:

```
numPersonas <- 50
resSampleo <- rpois( numPersonas, valorEsperado )
resSampleo
```

```
##  [1] 4 3 1 2 5 5 0 0 1 1 3 3 0 3 1 2 1 0 1 3 2 0 2
## [24] 0 1 0 4 6 0 2 1 2 3 1 2 2 2 4 2 1 2 0 2 2 3 2
## [47] 2 3 1 2
```

Variación técnica en cuentas de *RNA-seq* (Poisson) III

la media es igual a la varianza

```
valorEsperado
```

```
## [1] 2
```

```
var( resSampleo )
```

```
## [1] 2.05102
```

Variación técnica en cuentas de *RNA-seq* (Poisson) IV

El error de aproximación es de:

```
standarDev <- sqrt( valorEsperado )
standarDev / valorEsperado
```

```
## [1] 0.7071068
```

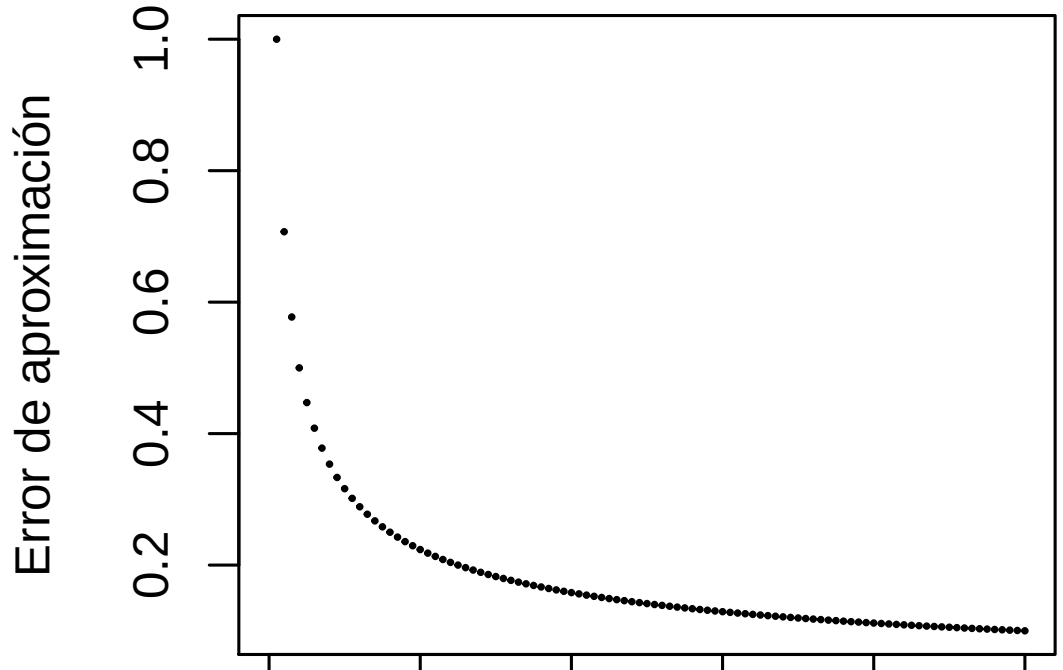
Que significa que los valores muestrados tienen una incertidumbre del 71%.
¿Qué sucede con el error de aproximación si el valor esperado incrementa? ¿Cómo podemos aumentar el valor esperado?

```
valorEsperado <- n * fracPelotasBlancas
```

Variación técnica en cuentas de *RNA-seq* (Poisson) IV

El error de aproximación depende de la media

```
n=seq(0, 1000, 10)
x <- obtenerValorEsperado( fracPelotasBlancas, n)
ea <- sqrt(x)/x
plot( n, ea, ylab="Error de aproximación", pch=19, cex=.2)
```



Variación técnica en cuentas de *RNA-seq* (Poisson) V

¿Cómo se traduce ésto a RNA-seq?

- Sampleo = pasos de sequenciacion (rtPCR)
- n = cobertura
- Varianza técnica puede ser estimado de los datos (varianza = media)
- Entre más cobertura, menos varianza técnica

Variación biológica en experimentos de RNA-seq

Klaus et al. The EMBO Journal. 2015