Laboratorio Nº3

Análisis de Datos

MVARGAS

Repasando lo ya visto

Modelo de Regresión Básico

- Mínimos cuadrados es una herramienta de estimación.
- ¿Cómo se usa para realizar inferencia?
- Para esto se desarrolla un modelo probabilístico de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Aquí ε_i se asume iid $N(0, \sigma^2)$.
- Note que $E[Y_i \mid X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- Note que $Var(Y_i \mid X_i = x_i) = \sigma^2$.
- La estimación por ML de β_0 y β_1 coincide con la estimación por OLS

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $E[Y \mid X = x] = \beta_0 + \beta_1 x$
- $Var(Y \mid X = x) = \sigma^2$

Interpretación de los coeficientes de regresión

Intercepto

- β_0 es el valor esperado del output cuando el input es 0

$$E[Y|X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

- Note que esto no siempre es de interés, por ejemplo cuando $X = 0$ es imposible o está fuera del rango de los datos (e.g. Si X corresponde a presión sanguínea, estatura, etc.)
- Considere que

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + a\beta_1 + \beta_1(X_i - a) + \varepsilon_i = \tilde{\beta}_0 + \beta_1(X_i - a) + \varepsilon_i$$

  Entonces, si desplazamos X en a unidades cambia el intercepto pero no la pendiente. menudo a se fija en $\bar{X}$ tal que el intercepto se interpreta como la respuesta esperada en el valor promedio de X.

Pendiente

- β_1 es el cambio esperado en el output cuando el input cambia en una unidad

$$E[Y \mid X = x + 1] - E[Y \mid X = x] = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x) = \beta_1$$

- Considere el impacto de cambiar las unidades (medición) de X

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + \frac{\beta_1}{a}(X_i a) + \varepsilon_i = \beta_0 + \tilde{\beta}_1 (X_i a) + \varepsilon_i$$

- Entonces, la multiplicación de X por un factor a resulta en que se divide el coeficiente por el mismo factor a.
- Ejemplo: X es la estatura en m e Y es el peso en kg. Entonces β_1 es kg/m. Convirtiendo X en cm implica multiplicar X por $100cm/m$. Para obtener β_1 en las unidades correctas, tenemos que dividir por $100cm/m$ y así se tendrán las unidades correctas.

$$Xm \times \frac{100cm}{m} = (100X)cm \quad \text{y} \quad \beta_1 \frac{kg}{m} \times \frac{1m}{100cm} = \left(\frac{\beta_1}{100}\right)\frac{kg}{cm}$$

Usando los coeficientes de regresión en una predicción

- Si queremos predecir el output dado un valor del input, digamos X, el modelo de regresión predice

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

Ejemplo: Base de datos `diamond` de la librería `UsingR`

Los datos son: * Precio de los diamantes (en dólares de Singapur) * Peso de los diamantes (en quilates) * Quilate = medida estándar del peso de un diamante = 0,2 g * Para obtener los datos hay que usar `library(UsingR); data(diamond)`

Gráfico

```r
library(UsingR)
data(diamond)
library(ggplot2)
g = ggplot(diamond, aes(x = carat, y = price))
g = g + xlab("Mass (carats)")
g = g + ylab("Price (SIN $)")
g = g + geom_point(size = 7, colour = "black", alpha=0.5)
g = g + geom_point(size = 5, colour = "blue", alpha=0.2)
g = g + geom_smooth(method = "lm", colour = "black")
g
```

Ajuste del modelo de regresión

```r
fit <- lm(price ~ carat, data = diamond)
coef(fit)
```

```
(Intercept)       carat
  -259.6259   3721.0249
```

- Se estima un aumento esperado de 3721.02 dólares de Singapur en el precio por un aumento de un quilate en el precio del diamante.

- El intercepto -259.63 corresponde al precio esperado de un diamante de 0 quilates.

Si se quiere información más detallada

```r
summary(fit)
```

```

Call:
lm(formula = price ~ carat, data = diamond)

Residuals:
    Min      1Q  Median      3Q     Max 
-85.159 -21.448  -0.869  18.972  79.370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
carat        3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778 
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

Obtención de un intercepto interpretable

Se puede escribir el modelo usando la desviación con respecto a la media $(X - \bar{X})$ como input.

```r
fit2 <- lm(price ~ I(carat - mean(carat)), data = diamond)
coef(fit2)
```

```
           (Intercept) I(carat - mean(carat)) 
              500.0833              3721.0249 
```

Entonces \$500.1 es el precio esperado para un diamante de peso promedio que en el caso de los datos corresponde a 0.2041667 quilates.

Cambio de escala

- Un incremento de 1 quilate es muy grande, ¿qué se esperaría si el peso aumenta 1/10 quilates?
- Se puede dividir el coeficiente por 10.
- Se espera un aumento de 372.102 dólares de Singapur en el precio por cada 1/10 quilates que aumenta el precio.
- Esto es lo mismo que cambiar la escala de X y ajustar la regresión

```r
fit3 <- lm(price ~ I(carat * 10), data = diamond)
coef(fit3)
```

```
  (Intercept) I(carat * 10) 
    -259.6259      372.1025 
```

Predicción del precio de un diamante

Supongamos que tenemos tres diamantes cuyos pesos son 0.16, 0,27 y 0,34 quilates. Estos serán los nuevos X aparte de los X que ya están en la base de datos. Entonces la estimación de su precio se obtiene de la siguiente forma:

```
newx <- c(0.16, 0.27, 0.34)
coef(fit)[1] + coef(fit)[2] * newx
```

```
[1]  335.7381  745.0508 1005.5225
```

```
predict(fit, newdata = data.frame(carat = newx))
```

```
        1         2         3 
 335.7381  745.0508 1005.5225 
```

```
newy <- coef(fit)[1] + newx + coef(fit)[2] * newx
```

Gráfico para interpretar la regresión

- Valores observados de los X de la base de datos $\rightarrow$ color azul
- Valores esperados de los X de la base de datos $\rightarrow$ color rojo
- Valores estimados de los X nuevos (los 3 diamantes de la parte anterior) $\rightarrow$ líneas rectas

```
data(diamond)
plot(diamond$carat, diamond$price,
     xlab = "Peso (quilates)",
     ylab = "Precio (dolares de Singapur)",
     bg = "royalblue",
     xlim=c(0.1, 0.4),
     col = "black", cex = 1.1, pch = 21,frame = FALSE)
abline(fit, lwd = 2)
points(diamond$carat, predict(fit), pch = 19, col = "red")
lines(c(0.16, 0.16, 0.1),
      c(200, coef(fit)[1] + coef(fit)[2] * 0.16,
      coef(fit)[1] + coef(fit)[2] * 0.16))
lines(c(0.27, 0.27, 0.1),
      c(200, coef(fit)[1] + coef(fit)[2] * 0.27,
        coef(fit)[1] + coef(fit)[2] * 0.27))
lines(c(0.34, 0.34, 0.1),
      c(200, coef(fit)[1] + coef(fit)[2] * 0.34,
        coef(fit)[1] + coef(fit)[2] * 0.34))
text(newx, rep(220, 3), labels = newx, pos = 4)
text(rep(0.12,3), round(newy, digits=2), labels =  round(newy, digits=2), pos = 3)
```

Regresión usando ANOVA

Usaremos la base de datos `mtcars` que viene en la librería `datasets`. Analizaremos cuál o cuáles variables nos permiten predecir la variable MPG (millas por galón) dado un conjunto de datos (rendimiento, peso, transmisión, etc) de varios modelos de automóviles.

Para ver las variables de esta base de datos hacemos lo siguiente:

```r
library(datasets) #This library provides free databases
data(mtcars) #The database I will use
str(mtcars) #str displays variables names and displays basic information
```

```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

ANOVA explica las fuentes de variabilidad, es decir que puede ayudar a determinar cuáles son las variables que tienen efectos significativos estadísticamente hablando. De acuerdo a Wikipedia: "En su forma más simple, ANOVA provee un test estadístico que permite determinar si las medias de varios grupos son iguales".

Ahora aplicamos ANOVA para determinar los efectos de todas las variables sobre MPG en el contexto de un modelo lineal.

```r
analysis <- aov(mpg ~ ., data = mtcars) #I run ANOVA
summary(analysis) #this returns a summary containing relevant statistics
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
cyl          1  817.7   817.7 116.425 5.03e-10 ***
disp         1   37.6    37.6   5.353  0.03091 *  
hp           1    9.4     9.4   1.334  0.26103    
drat         1   16.5    16.5   2.345  0.14064    
wt           1   77.5    77.5  11.031  0.00324 ** 
qsec         1    3.9     3.9   0.562  0.46166    
vs           1    0.1     0.1   0.018  0.89317    
am           1   14.5    14.5   2.061  0.16586    
gear         1    1.0     1.0   0.138  0.71365    
carb         1    0.4     0.4   0.058  0.81218    
Residuals   21  147.5     7.0                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo 1

Estimaremos el siguiente modelo considerando los datos de la tabla ANOVA:

$$MPG_i = \beta_0 + \beta_1 CYL_i + \beta_2 DISP_i + \beta_3 WT_i + \beta_4 AM_i + \varepsilon_i$$

```r
fit1 <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(fit1)
```

```
Call:
lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-4.318 -1.362 -0.479  1.354  6.059 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
cyl         -1.784173   0.618192  -2.886  0.00758 ** 
disp         0.007404   0.012081   0.613  0.54509    
wt          -3.583425   1.186504  -3.020  0.00547 ** 
am           0.129066   1.321512   0.098  0.92292    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.642 on 27 degrees of freedom
Multiple R-squared:  0.8327,    Adjusted R-squared:  0.8079 
F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

Modelo 2

Estimaremos el siguiente modelo considerando la significancia de las variables:

$$MPG_i = \beta_0 + \beta_1 CYL_i + \beta_2 WT_i + \beta_3 AM_i + \varepsilon_i$$

```r
fit2 <- lm(mpg ~ cyl + wt + am, data = mtcars)
summary(fit2)
```

```
Call:
lm(formula = mpg ~ cyl + wt + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1735 -1.5340 -0.5386  1.5864  6.0812 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
cyl          -1.5102     0.4223  -3.576  0.00129 ** 
wt           -3.1251     0.9109  -3.431  0.00189 ** 
am            0.1765     1.3045   0.135  0.89334    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.612 on 28 degrees of freedom
Multiple R-squared:  0.8303,    Adjusted R-squared:  0.8122 
F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Ejercicio

Escriba un informe de no más de 2 páginas junto a su grupo de trabajo (los mismos del proyecto de curso) en el que se responda claramente:

- ¿Cuáles son las variables más importantes para explicar la variable MPG en distintos modelos de automóviles?
- ¿Cuál es el efecto de las variables más importantes sobre la varible MPG? De una interpretación simple de cada una de las variables (e.g. si la variable X_j aumenta en a unidades se espera que la variable Y aumente/disminuya b unidades)
- ¿Cuál de los dos modelos de la parte anterior es mejor y por qué?
- ¿Se puede decir que la variable AM (tipo de transmisión) explica la variable MPG en distintos modelos de automóviles?
- Comente los alcances y limitaciones del modelo en base a los datos disponibles y los supuestos de OLS

Indicaciones:

- Use argumentos estadísticos y argumentos teóricos en base a prensa especializada
- Presente sus resultados en un lenguaje simple y formal