Biostatistics Exercises

www.nataliejulian.com

Propensity Score

La base de datos `prop_score.txt` contiene información sobre una campaña realizada para promocionar la compra de un artículo. La variable sent contiene información referente a si a la persona se le envió un email informando sobre la promoción (1=Sí y 0=No) y la variable bought contiene información referente a si la persona compró o no el artículo (1=Sí y 0=No).

Considere como variable respuesta, Y, la variable bought, como variable de asignación de tratamiento, Z, la variable sent y como covariable, X, la variable Income (valor estandarizado del ingreso).

a) Utilice regresión logística y estime $E(Y|X, Z = 0)$, $E(Y|X, Z = 1)$ y los propensity score para cada individuo. Con las cantidades anteriores, grafique $(X, \widehat{E(Y|X)})$ considerando el propensity score.

Hint: Si Z es una partición del espacio muestral, entonces:

$$E(Y|X) = \sum_z E(Y|X, Z = z)P(Z = z|X)$$

```
> library(readr)
> prop_score <- read_table2("prop_score.txt")
> names(prop_score)<-c("income","sent","bought")

> table(prop_score$bought)  #No hay ninguna categoría mucho más grande que la otra

  0   1
408 593
```

La variable de asignación de tratamientos es Z si se le envió o no un e-mail informando sobre la promoción. Es necesario estimar para cada una de las observaciones, el *propensity score* es decir, la probabilidad de que dicho cliente fuera "tratado" (i.e se le haya enviado un email) $P(Z = 1|X)$:

```
> prop_score$sent2<-factor(prop_score$sent)  #Se definen como factores
> prop_score$bought2<-factor(prop_score$bought)
> prop.score <- glm(sent2 ~ income,family = binomial,data = prop_score)
> #probabilidad que le sea aplicado tratamiento dada la covariable:
> PZ1 <- exp(coef(prop.score)[1]+
+                     coef(prop.score)[2]*prop_score$income)/(1+
+ exp(coef(prop.score)[1]+coef(prop.score)[2]*prop_score$income))
> #probabilidad que no sea tratamiento, i.e sea control dada la covariable
> PZ0<-1-PZ1
```

Ahora necesitamos obtener $E(Y|X, Z = 0)$, es decir dado que fue control, entregar una estimación de la probabilidad de que compre el producto

```
> library(dplyr)
> datacontrol<-prop_score%>%filter(sent=="0")
> modcontrol<-glm(bought2 ~ income,family = binomial,
+                  data = datacontrol)
> coef(modcontrol)

(Intercept)      income
   1.054439    1.792506

> #Probabilidades de que compren dada la covariable para el grupo control
> PYcontrol<-exp(coef(modcontrol)[1]+
+                        coef(modcontrol)[2]*prop_score$income)/(1+
+ exp(coef(modcontrol)[1]+ coef(modcontrol)[2]*prop_score$income))
>
```

Ahora necesitamos obtener $E(Y|X, Z = 1)$, es decir dado que fue tratado, entregar una estimación de la probabilidad de que compre el producto:

```
> datatrat<-prop_score%>%filter(sent=="1")
> modtrat<-glm(bought2 ~ income,family = binomial,
+              data = datatrat)
> coef(modtrat)

(Intercept)      income
 -0.2181215   5.9575375

> #Probabilidades de que compren dada la covariable para el grupo tratamiento
> PYtrat<-exp(coef(modtrat)[1]+
+                      coef(modtrat)[2]*prop_score$income)/(1+exp(coef(modtrat)[1]+
+ coef(modtrat)[2]*prop_score$income))
>
```

Ya que tenemos $P(Z = z|X)$ (probabilidad de que el individuo en base a sus características de Income haya sido tratamiento o control), $E(Y|X, Z = 0)$ que dado que Y es compra o no compra (es decir, es una variable de naturaleza Bernoulli) consiste en la probabilidad de que compre el producto dado X y dado $Z = 0$ (es decir, que fue control) y asimismo, ya quemos calculado $E(Y|X, Z = 1)$ i.e la probabilidad de que compre el producto dado X y dado $Z = 1$ (es decir, que fue tratado), podemos obtener $E(Y|X)$ utilizando el hint como sigue:

```
> #E(Y|X,Z=1)*P(Z=1)+E(Y|X,Z=0)*P(Z=0) para cada observacion Y bernoulli
>
> EYX<-PYtrat*PZ1+PYcontrol*PZ0
> summary(EYX)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02029 0.29877 0.46321 0.59277 0.97699 1.00000

>
```

Gráfico ordenado por Income:

```
> attach(prop_score)
> #Grafico:
>
> plot(income[order(income)], EYX[order(income)],
+      type="o",
+      xlab="Ingreso estandarizado",
+      ylab="Probabilidad",
+      main="Probabilidad vía propensity score de que la persona compre el producto"
+      pch=19,
+      col="lightsteelblue")
>
```

b) Sin considerar la descomposición anterior, estime $E(Y|X)$ mediante regresión logística y grafique $(X, \widehat{E(Y|X)})$. Discuta la ventaja de utilizar propensity scores para la estimación de $E(Y|X)$.

Omitiendo la información de si se le aplicó o no tratamiento a los clientes, se puede establecer una regresión logística usual:

```
> logist<-glm(bought2 ~ income,family = binomial,data = prop_score)
> plot(income[order(income)], logist$fitted.values[order(income)],
+      type="o",
+      xlab="Ingreso estandarizado",
+      ylab="Probabilidad",
+      main="Probabilidad de que la persona compre el producto",
```

```
+      pch=19,
+      col="lightsteelblue")
```

Es posible observar que al no considerar el factor tratamiento, es posible llegar a conclusiones muy diferentes, sobretodo en torno a la probabilidad 0.5, donde al omitir información de Z se asume la misma velocidad hacia la probabilidad 1 de adquirir el producto que a la probabilidad 0 de adquirir el producto (es decir, utilizando una regresión logística directamente se establece simetría en el valor de probabilidad 0.5, la que considerando mayor información claramente no existiría). Es evidente que no puede asumirse un mismo comportamiento en términos de comprar o no el producto para aquellos clientes que recibieron un email y para aquellos que no, dicha información **debe** considerarse en el modelamiento, y es precisamente esa la idea del propensity score, manejar ese sesgo en los resultados, producidos por la variable tratamiento aplicado a las unidades experimentales.