Lista 9: Minimos cuadrados ponderados y generalizados

Claudia Meneses

4/1/2020

Con base en los datos sobre gasto en educación para el año de 1975 en USA, ajustamos un modelo de regresión lineal múltiple.

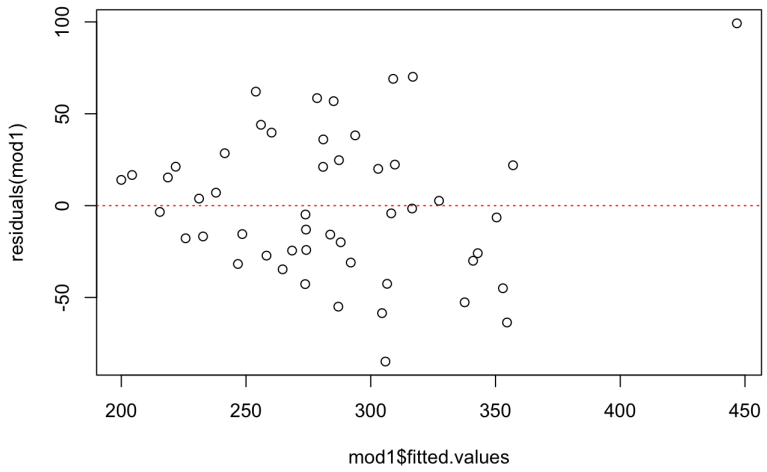
Con excepción de x_3 todos los coeficientes son significativos. La R^2 no es muy buena esto puede deberse a que hacen falta regresores.

```
mod1 <- lm(y~x1+x2+x3, data=datos)
summary(mod1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.878 -26.878  -3.827  22.246  99.243
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.566e+02  1.232e+02  -4.518 4.34e-05 ***
## x1           7.239e-02  1.160e-02   6.239 1.27e-07 ***
## x2           1.552e+00  3.147e-01   4.932 1.10e-05 ***
## x3          -4.269e-03  5.139e-02  -0.083    0.934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.47 on 46 degrees of freedom
## Multiple R-squared:  0.5913, Adjusted R-squared:  0.5647
## F-statistic: 22.19 on 3 and 46 DF,  p-value: 4.945e-09
```

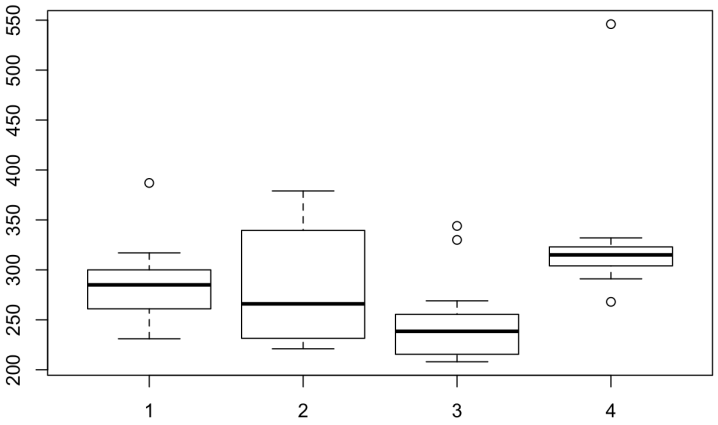
La gráfica de los residuales contra $\hat{y}$ parece indicar que entre mayor es $\hat{y}$ mayor es la varianza de los residuales. Además se puede observar un outlier en la parte superior derecha.

```
plot(mod1$fitted.values,residuals(mod1))
abline(h=0, lty="dotted", col="red")
```

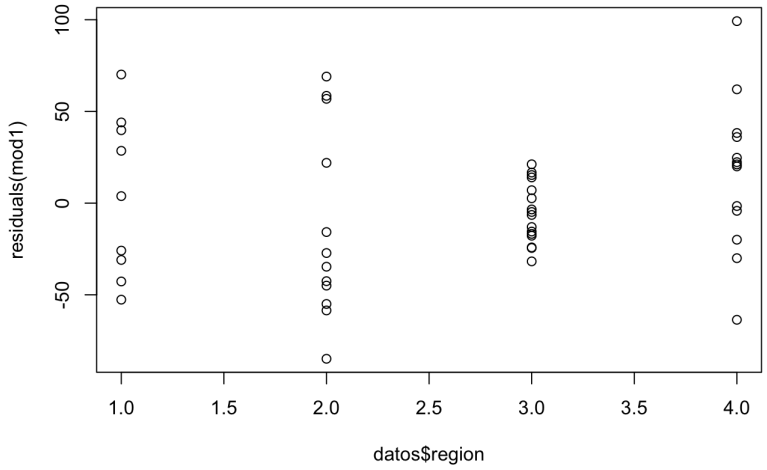


La gráfica de y (gasto per cápita para la educación) de acuerdo a la región parece indicar que la varianza de y está muy relacionada con la región.

```
boxplot(datos$y~as.factor(datos$region))
```




```
plot(residuals(mod1)~datos$region)
```



Al parecer la observación 49 parece ser un outlier influencial. Lo removemos para tener un mejor modelo.

```
mod2 <- lm(y~x1+x2+x3, data=datos[-49,])
summary(mod2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = datos[-49, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.128 -22.154  -7.542  22.542  80.890
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -277.57731  132.42286  -2.096 0.041724 *
## x1             0.04829    0.01215   3.976 0.000252 ***
## x2             0.88693    0.33114   2.678 0.010291 *
## x3             0.06679    0.04934   1.354 0.182591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.81 on 45 degrees of freedom
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4631
## F-statistic:  14.8 on 3 and 45 DF,  p-value: 7.653e-07
```

Calculamos la varianza por región a partir de los residuales para este modelo. Para cada región la varianza es muy diferente.

```
res <- residuals(mod2)
res_1 <- res[datos$region==1]
n1 <- length(res_1)
var_1 <- var(res_1)
var_1
```

```
## [1] 1629.759
```

```
var_1 <- rep(var_1,n1)

res_2 <- res[datos$region==2]
n2 <- length(res_2)
var_2 <- var(res_2)
var_2
```

```
## [1] 2640.893
```

```
var_2 <- rep(var_2,n2)

res_3 <- res[datos$region==3]
n3 <- length(res_3)
var_3 <- var(res_3)
var_3
```

```
## [1] 185.8242
```

```
var_3 <- rep(var_3, n3)

res_4 <- res[datos$region==4][-length(res[datos$region==4])]
n4 <- 13
var_4 <- na.omit(var(res_4))
var_4
```

```
## [1] 810.2099
```

```
var_4 <- rep(var_4, n4)
```

Calculamos las ponderaciones w_i para la estimación de mínimos cuadrados ponderados (WLS).

```
vars <- c(var_1,var_2,var_3,var_4)
wi <- sqrt(1/vars)
```

Calculamos la estimación de mínimos cuadrados ponderados usando los w_i. El $Adj - R^2$ mejoró bastante, sin embargo sólo dos coeficientes son significantes.

```
R=diag(x=wi)
w <- (solve(R)%*%datos$y)[-49]
Z <- (solve(R)%*%model.matrix(mod1))[-49,]
mcp <- lm(w~0+Z[,1]+Z[,2]+Z[,3]+Z[,4])
summary(mcp)
```

```
##
## Call:
## lm(formula = w ~ 0 + Z[, 1] + Z[, 2] + Z[, 3] + Z[, 4])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4210.2  -514.6  -221.9   634.9  4185.1
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## Z[, 1] -238.75993  194.80607  -1.226   0.2267
## Z[, 2]    0.03598    0.01755   2.050   0.0462 *
## Z[, 3]    0.89438    0.48428   1.847   0.0714 .
## Z[, 4]    0.09504    0.06239   1.523   0.1347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1543 on 45 degrees of freedom
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.977
## F-statistic: 520.6 on 4 and 45 DF,  p-value: < 2.2e-16
```