Regresión cuadrática - Econometría I

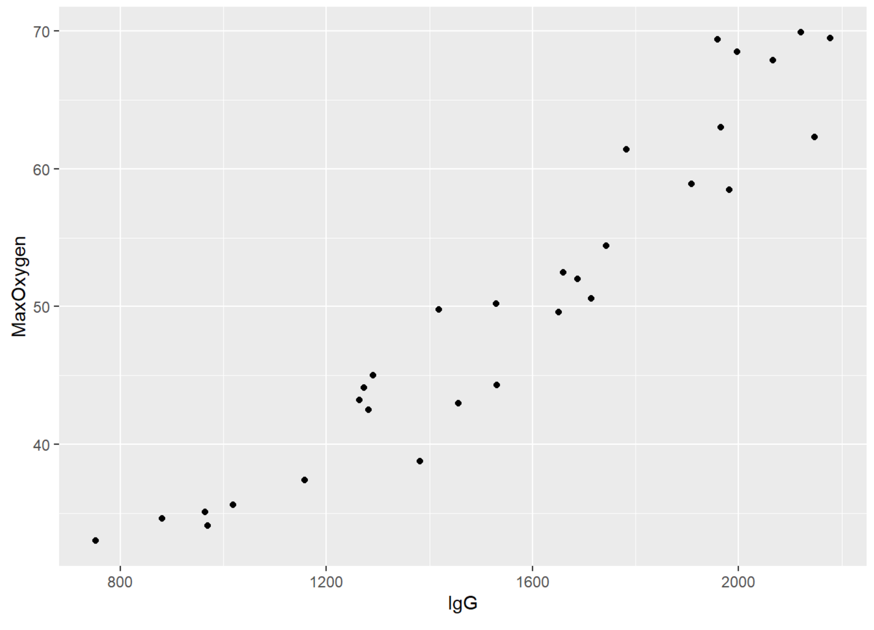
Agustin Riquelme y Heriberto Espino

La base que se usará para este trabajo será la del sistema inmune. como podemos observar, que existen dos distintas variables, una corresponde a la máxima oxigenación en sangre (MaxOxygen) y otra corresponde al número de Inmunoglobina tipo G. Para este caso la máxima oxigenación se usará como variable de respuesta y el número de Inmunoglobina G como variable predictora

```
## # A tibble: 6 × 3
##   sujeto   IgG MaxOxygen
##    <dbl> <dbl>     <dbl>
## 1      1   881      34.6
## 2      2  1290      45
## 3      3  2147      62.3
## 4      4  1909      58.9
## 5      5  1282      42.5
## 6      6  1530      44.3
```

Podemos observar con nuestro gráfico de puntos que el modelo parece no ser lineal por completo, sino se asemeja más a una función exponencial que va creciendo más conforme aumenta el nivel de Inmunoglobina tipo G en sangre, de todas maneras se va a desarrollar un modelo líneal para ver las diferencias entre sí

```
ggplot(SI, aes(x=IgG, y=MaxOxygen)) +
  geom_point()
```



1. Modelo Lineal Simple

Primero se realizará un modelo de regresión lineal simple, donde se usarán las variables de la misma manera que hemos trabajado todo el tiempo, podemos observar en dicho modelo que la variable es significativa, tanto la beta 1 como la 0 son variables que con un nivel significativas en un nivel de 0.05 y el modelo en general igual obtiene dicha calificación, pues su R^2 tanto normal como ajustada tiene un valor alto, diciendo que el 91% de nuestros datos se ajustan al modelo de manera correcta, además el error estándar es relativamente pequeño por lo que podemos decir que es un buen modelo pero para afirmar que es el mejor para esta base, será necesario compararlo con los otros modelos

```
mod_lin <- lm(y ~ x)
summary(mod_lin)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9491 -2.2064  0.1563  2.5474  7.5940
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.290506   2.642545   2.759   0.0101 *
## x           0.027828   0.001642  16.947 2.97e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.638 on 28 degrees of freedom
## Multiple R-squared:  0.9112, Adjusted R-squared:  0.908
## F-statistic: 287.2 on 1 and 28 DF,  p-value: 2.973e-16
```

A través de la prueba de normalidad Shapiro-Wilks podemos observar que se tiene un p valor muy alto, es decir, que aceptamos la hipótesis nula de que los residuos se comportan normalmente con un nivel de significacia de 0.05. Además con el test de Durbin-Watson podemos observar que de igual manera el p valor es muy alto, por lo que aceptamos la hipótesis nula y decimos que no existe correlación entre los residuos

```
shapiro.test(mod_lin$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_lin$residuals
## W = 0.98777, p-value = 0.9747
```

```
dwtest(mod_lin)
```

```
##
##  Durbin-Watson test
##
## data:  mod_lin
## DW = 2.6006, p-value = 0.9511
## alternative hypothesis: true autocorrelation is greater than 0
```

Y finalmente, hacemos sacamos el valor del valor de Aikake para poder comparar qué modelo es mejor, en este caso tenemos un valor de 166.5 que resulta alto

```
AIC(mod_lin)
```

```
## [1] 166.5496
```

2. Modelo Cuadrático Simple

Ahora realizaremos un modelo cuadrático para ver los cambios que se pueden observar en comparación con el modelo de regresión lineal simple. Principalmente podemos observar que el valor de nuestra R^2 aumentó a 92%, es decir, que 1% más de los datos se ajusta a nuestro modelo. De igual manera el error residual estándar es más pequeño que el anterior y los p valores de tanto beta 1 como beta 0, son mucho más significativos que el modelo pasado

```
mod_cuad <- lm(y ~ I(x^2))
summary(mod_cuad)
```

```
##
## Call:
## lm(formula = y ~ I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9227 -1.6685  0.0318  1.7592  7.2806
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.681e+01  1.407e+00   19.06   <2e-16 ***
## I(x^2)      9.202e-06  4.904e-07   18.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.313 on 28 degrees of freedom
## Multiple R-squared:  0.9263, Adjusted R-squared:  0.9237
## F-statistic: 352.1 on 1 and 28 DF,  p-value: < 2.2e-16
```

Además podemos observar que con un valor más significativo, los residuos se comportan de manera normal con la prueba Shapiro-Wilk. Sin embargo, para el caso del test de Durbin-Watson, sí aceptamos la hipótesis nula pero con un nivel menos significativo

```
shapiro.test(mod_cuad$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_cuad$residuals
## W = 0.99024, p-value = 0.9922
```

```
dwtest(mod_cuad)
```

```
##
##  Durbin-Watson test
##
## data:  mod_cuad
## DW = 2.335, p-value = 0.8121
## alternative hypothesis: true autocorrelation is greater than 0
```

Con todo esto, comparado con el modelo anterior, el valor de Aikake resulta más bajo

```
AIC(mod_cuad)
```

```
## [1] 160.9323
```

3. Modelo Cuadrático Completo

Para el modelo cuadrático completo podemos observar que el modelo resulta más significativo a pesar de ajustarse a los datos de manera similar al modelo anterior, sin embargo las betas resultan ser menos significativas e incluso nuestra beta 1 puede ser removida del modelo al no aportar mucha significancia

```
mod_cuadcom <- lm(y ~ x + I(x^2))
summary(mod_cuadcom)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9246 -1.6683  0.0317  1.7580  7.2805
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.683e+01  8.640e+00   3.105  0.00444 **
## x           -2.894e-05  1.191e-02  -0.002  0.99808
## I(x^2)       9.211e-06  3.907e-06   2.358  0.02588 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.373 on 27 degrees of freedom
## Multiple R-squared:  0.9263, Adjusted R-squared:  0.9209
## F-statistic: 169.8 on 2 and 27 DF,  p-value: 5.102e-16
```

Comparado con el modelo anterior, los residuos se comportan de manera similar, tanto en la prueba de normalidad como en la correlación entre ellos, así podemos afirmar que se comportan de manera normal y los residuos no tienen correlación entre sí

```
shapiro.test(mod_cuadcom$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_cuadcom$residuals
## W = 0.99024, p-value = 0.9922
```

```
dwtest(mod_cuadcom)
```

```
##
##  Durbin-Watson test
##
## data:  mod_cuadcom
## DW = 2.3347, p-value = 0.8255
## alternative hypothesis: true autocorrelation is greater than 0
```

Este modelo tiene un valor de Aikake más alto que el anterior, sin embargo, resulta de igual manera más bajo que el modelo de regresión lineal

```
AIC(mod_cuadcom)
```

```
## [1] 162.9323
```

4. Modelo Polinomial

Ahora para el caso del modelo polinomial, podemos observar que se ajustan los datos de manera similar al modelo anterior al respresentar al 92% de los datos con el modelo, sin embargo podemos observar que la beta 3 puede ser removida del modelo al no resultar tan significativa. Y respecto al error estándar en los residuos, es mayor que en los modelos cuadráticos pero menor que en lineal simple

```
mod_pol <- lm(y ~ poly(x,3))
summary(mod_pol)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2862 -1.7814 -0.0516  2.0736  6.9326
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.6367     0.6209  81.556  < 2e-16 ***
## poly(x, 3)1  61.6508     3.4007  18.129 2.84e-16 ***
## poly(x, 3)2   7.9545     3.4007   2.339   0.0273 *
## poly(x, 3)3  -2.5655     3.4007  -0.754   0.4574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.401 on 26 degrees of freedom
## Multiple R-squared:  0.9279, Adjusted R-squared:  0.9196
## F-statistic: 111.6 on 3 and 26 DF,  p-value: 5.735e-15
```

A su vez podemos decir que los errores se comportan de manera normal al no obtener un p valor significativo en la prueba de Shapiro-Wilks y afirmar que no existe una correlación entre los residuos por el test de Durbin-Watson

```
shapiro.test(mod_pol$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_pol$residuals
## W = 0.98662, p-value = 0.9615
```

```
dwtest(mod_pol)
```

```
##
##  Durbin-Watson test
##
## data:  mod_pol
## DW = 2.3527, p-value = 0.8377
## alternative hypothesis: true autocorrelation is greater than 0
```

Finalmente el valor de Aikake de este último modelo es más grande que los anteriores pero aún así menor que el caso del modelo lineal. Sin embargo, este modelo podría resultar mejor ya que su criterio a pesar de ser más bajo que el del modelo lineal, es mucho mejor que los cuadráticos, y sabiendo que dicho criterio penaliza a aquellos modelos complejos a favor de los sencillos para evitar un sobreajuste, podemos decir que este modelo resulta más significativo y tiene mucha mejor confianza.

```
AIC(mod_pol)
```

```
## [1] 164.2827
```