# Tarea 3

## Luis Federico Puente Peña

## Parte teórica

### PARTE 1.

### 1. Plantear el problema de regresión lineal como un problema de mínimos cuadrados, encontrar el vector beta que resuelva

Sea entonces

Derivando con respecto a cada uno de los elementos de obtenemos: Entonces, si existe se tiene que

### ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos?

El planteamiento original del problema es el siguiente:

El cual requiere que sea lineal en los parámetros, es decir, cada unas de los elementos de está elevado a la potencia 1. Sin embargo, puede no ser lineal en las variables, es decir, cada una las variables explicativas puede ser una una transformación .

### ¿Podríamos usarlo para ajustar polinomios (ej )?

De ahí que sea posible ajustar la variable a un polinomio, por ejemplo

Podemos ajustar polinomios del modo y mayores sin ningún problema utilizando el resultado de regresión lineal para las 's. Lo que es importante notar es que aunque la regresión de polinomio ajusta un modelo no-lineal sobre los datos, el problema de estimación estadística continúa siendo lineal (es lineal en las 's aunque no en las variables x) lo que es consecuencia directa de que la función E(Y|X) es lineal en los parámetros beta estimados.

### 2. Argumentar la relación entre la solución encontrada y un problema de proyección en subespacios vectoriales de álgebra lineal ¿Cuál es la relación particular con el teorema de Pitágoras?

Cuando hay una relación lineal entre dos variables, la varianza de la variable dependiente se puede descomponer en dos varianzas: la de los pronósticos, debido a la relación que la variable dependiente guarda con la variable independiente, y la de los errores o residuos. Esta relación se cumple tanto para la Regresión Lineal Simple como para la Múltiple. Esta descomposición de la varianza de la variable dependiente en dos varianzas es el "Teorema de Pitágoras" del Análisis de Regresión Lineal que, para efectos del modelo anterior, la varianza de las puntuaciones observadas es igual a la varianza de las puntuaciones estimadas más la varianza de los residuos.

### 3. ¿Que logramos al agregar una columna de unos en la matriz X? es decir, definir mejor

Al agregar la columna de 1s se plantear el problema de la siguiente forma:

Por lo que se agrega el estimador a la forma funcional de Y, esto genera que exista una constante para estimar Y, por lo que, a diferencia del caso anterior, la función puede no pasar por el origen. Es ayuda a reducir los errores de estimación cuando hay existe una constante (que no depende de las variables) para estimar el modelo.

### 4. Plantear el problema de regresión ahora como un problema de estadística

### donde los errores son no correlacionados con distribución

Si a la ecuación anterior se le aplica la esperanza tenemos: ya que o dicho de otra forma:

### 5. ¿Cual es la función de verosimilitud del problema anteriror? Hint: empiecen por excribir el problema como

### Sea

### con

### entonces

### además (considerando independencia)

### con la matriz identidad. Y concluyan entonces que

### Escriban entonces la verosimilitud como

### 6. Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados. La función log de máxima verosimilitud es:

El siguiente paso es derivar respecto a cada una de las :

Que es igual a cero solo si

Esto se satisface si:

### 7. Investiga el contenido del Teorema de Gauss-Markov sobre mínimos cuadrados.

El Teorema de Gauss-Márkov establece que en un modelo lineal general (MLG) en el que se cumplan los siguientes supuestos: - Correcta especificación: el MLG ha de ser una combinación lineal de los parámetros y no necesariamente de las variables: - Muestreo aleatorio simple: la muestra de observaciones del vector es una muestra aleatoria simple y, por lo tanto, el vector es independiente del vector   
- Esperanza condicionada de los errores nula: - Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N - Homocedasticidad: la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones:

El estimador mínimo cuadrático ordinario (MCO) de es el estimador lineal e insesgado óptimo, es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.

## Parte aplicada

### Para esta parte pueden usar la base de datos diamonds que sugirieron, aunque hay puntos adicionales si usan alguna base original interesante.

### Cargar la base que se encuentra en el paquete ggplot2. Los comandos que pueden usar para cargar la base diamonds a su ambiente de trabajo en R son:

if(!require(ggplot2)) install.packages("ggplot2",repos = "http://cran.us.r-project.org")

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.1

library(foreign)  
wage1 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")  
head(wage1)

## wage educ exper tenure nonwhite female married numdep smsa northcen  
## 1 3.10 11 2 0 0 1 0 2 1 0  
## 2 3.24 12 22 2 0 1 1 3 1 0  
## 3 3.00 11 2 0 0 0 0 2 0 0  
## 4 6.00 8 44 28 0 0 1 0 1 0  
## 5 5.30 12 7 2 0 0 1 1 0 0  
## 6 8.75 16 9 8 0 0 1 0 1 0  
## south west construc ndurman trcommpu trade services profserv profocc  
## 1 0 1 0 0 0 0 0 0 0  
## 2 0 1 0 0 0 0 1 0 0  
## 3 0 1 0 0 0 1 0 0 0  
## 4 0 1 0 0 0 0 0 0 0  
## 5 0 1 0 0 0 0 0 0 0  
## 6 0 1 0 0 0 0 0 1 1  
## clerocc servocc lwage expersq tenursq  
## 1 0 0 1.131402 4 0  
## 2 0 1 1.175573 484 4  
## 3 0 0 1.098612 4 0  
## 4 1 0 1.791759 1936 784  
## 5 0 0 1.667707 49 4  
## 6 0 0 2.169054 81 64

### Posteriormente deben hacer una regresión lineal. Su objetivo es explicar la variable price usando las demás variables. Noten que algunas variables no son numéricas, por lo que no pueden incluirse en un análisis crudo de regresión lineal. Para este proyecto NO es necesario saber transformar las variables no numéricas para poder usarlas en la regresión; hacerlo es optativo, de hecho, las paqueterías lo hacen por ustedes pero deben ser cuidadosos. Pueden usar la función lm de R para su análisis de regresión.

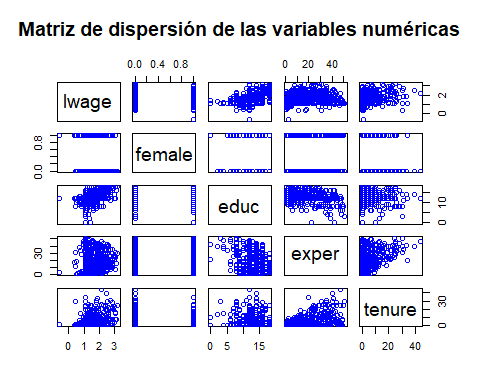
# replace "female" with logical variable  
#wage1$female <- as.logical(wage1$female)  
table(wage1$female)

##   
## 0 1   
## 274 252

Salarios <- wage1[c("lwage","female","educ","exper","tenure")]

Obtenemos una matriz de dispersión para formarnos una apreciación inicial de las relaciones entre las variables.

pairs(Salarios, col= "blue", main="Matriz de dispersión de las variables numéricas")

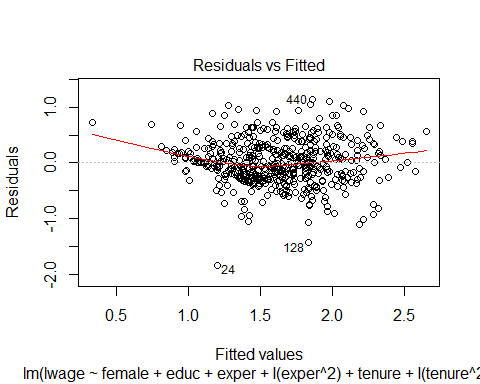


Gracias a la matriz de las variables en pares podemos discernir de forma rápida una relación positiva entre el salario (medido en logaritmos,lwage) y la variable de educación (educ), aunque la varianza aumenta conforme aumenta esta variable. También podemos observar que hay relación entre el salario y otras variables como ser hombre.

#Realizamos la regresión lineal usando todas las variables numéricas  
modelo1 = lm(lwage ~ female+educ+exper+I(exper^2)+tenure+I(tenure^2), data=Salarios)  
summary(modelo1)

##   
## Call:  
## lm(formula = lwage ~ female + educ + exper + I(exper^2) + tenure +   
## I(tenure^2), data = Salarios)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.83160 -0.25659 -0.02126 0.25500 1.13370   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4166909 0.0989279 4.212 2.98e-05 \*\*\*  
## female -0.2965110 0.0358055 -8.281 1.04e-15 \*\*\*  
## educ 0.0801967 0.0067573 11.868 < 2e-16 \*\*\*  
## exper 0.0294324 0.0049752 5.916 6.00e-09 \*\*\*  
## I(exper^2) -0.0005827 0.0001073 -5.431 8.65e-08 \*\*\*  
## tenure 0.0317139 0.0068452 4.633 4.56e-06 \*\*\*  
## I(tenure^2) -0.0005852 0.0002347 -2.493 0.013 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3998 on 519 degrees of freedom  
## Multiple R-squared: 0.4408, Adjusted R-squared: 0.4343   
## F-statistic: 68.18 on 6 and 519 DF, p-value: < 2.2e-16

plot(modelo1, which =1)



### 1. ¿Qué tan bueno fue el ajuste?

Del "summary" de la regresión podemos observar inicialmente la información de los residuales, que son todas las diferencias entre el salario estimado y el salario real; una media lo más cercana a cero es lo que querríamos ver aquí, pues indicaría que el modelo es muy certero. En la gráfica vemos los valores predichos contra los residuales, en un modelo perfecto, la media de los residuales sería 0 por lo que la línea roja iría perfectamente a lo largo de la línea que representa el 0. En seguida podemos apreciar los coeficientes de la intercepción y de cada una de las variables, usados para calcular las predicciones. De los valores "Pr(>|t|)" podemos ver que es importante incluir estas varibles en el modelo, ya que todas resultaron significativas.

### 2. ¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cúal fue el valor de $ ^2 $ que ajustó su modelo y que relación tienen con la calidad del ajuste?

Al final del summary podemos apreciar la R cuadrada, que es la cantidad de variabilidad en lo que estas prediciendo que es explicado por el modelo, en este caso podríamos decir que un 44.08 % del salario es explicado por las variables en nuestro modelo. Es una buena medida de la calidad del ajuste.

### 3. ¿Cual es el angulo entre y estimada? Hint: usen la cuadrada y el arcocoseno

angulo <- acos(sqrt(.4408))  
angulo \* 180/pi

## [1] 48.39989

### 4. Definan una funcion que calcule la logverosimilitud de unos parámetros y .

lm.loss <- function(par) {  
x.0 <- par[1]  
x.female <- par[2]  
x.educ <- par[3]  
x.exper <- par[4]  
x.exper2 <- par[5]  
x.tenure <- par[6]  
x.tenure2 <- par[7]  
err.sigma <- par[8]  
  
likelihoods <- dnorm(Salarios$lwage, mean = x.0 + x.female%\*%Salarios$female + x.educ%\*%Salarios$educ + x.exper%\*%Salarios$exper+x.exper2%\*%I(Salarios$exper^2)+x.tenure%\*%Salarios$tenure+x.tenure2%\*%I(Salarios$tenure^2), sd = err.sigma)  
log.likelihoods <- log(likelihoods)  
  
sum.log.like <- sum(log.likelihoods)  
  
return(sum.log.like)  
   
}  
  
  
valor <- lm.loss(c(0.4166909,-0.2965110,0.0801967,0.0294324,-0.0005827,0.0317139,-0.0005852,1))  
valor

## [1] -524.837

### 5. Utilicen la función optim de R para numericamente el máximo de la función de verosimilitud. Si lo hacen correctamente, su solución debe coincidir con la del método lm.

parameter.fits <- optim(par = c(0.4166909,-0.2965110,0.0801967,0.0294324,-0.0005827,0.0317139,-0.0005852,1),fn = lm.loss)  
   
  
parameter.fits

## $par  
## [1] -0.097053723 -1.053877814 -0.261099807 2.170436127 -0.006077952  
## [6] -0.733747319 0.019080324 2.285021307  
##   
## $value  
## [1] -67465.62  
##   
## $counts  
## function gradient   
## 501 NA   
##   
## $convergence  
## [1] 1  
##   
## $message  
## NULL