

Práctica 3. Análisis de datos y regresión

Estadística


Grado en Ingeniería Informática

Índice

1. Análisis descriptivo	2
2. La recta de regresión	5
2.1. El modelo lineal	5
2.2. Estimación por mínimos cuadrados	6
2.3. Algunos coeficientes de interés	8

1. Análisis descriptivo

La **Estadística Descriptiva** es un conjunto de técnicas numéricas y gráficas para describir y analizar un grupo de datos, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen. En el Tema 1 se han introducido algunas técnicas descriptivas básicas, como la construcción de tablas de frecuencias, la elaboración de gráficas y las principales medidas descriptivas de centralización, dispersión y forma que permitirán realizar la descripción de datos. Además, en la Práctica 2 hemos aplicado todas estas técnicas de Estadística Descriptiva tanto para variables cualitativas como cuantitativas.

Ejercicio. Descarga del Campus Virtual el archivo `DatosGapminder.csv`. Este archivo ha sido obtenido de la web <http://www.gapminder.org/>. Importa los datos a  y comprueba que todo ha funcionado correctamente con el comando `head`. Se pide:

1. Resumen gráfico de la variable PCs.
2. Estudia la relación entre GDP y PCs a través de la covarianza y del coeficiente de correlación.
3. Varianza y desviación típica de la variable PCs.
4. Cuantiles de orden 0.3, 0.6 y 0.9 de la variable PCs.
5. Tipificar la variable PCs.
6. Realizar el mismo análisis descrito en los pasos anteriores para la variable GDP.

Comenzaremos el análisis con la lectura de datos:

```
> datos<-read.csv(file="DatosGapminder.csv")
> head(datos)
```

```
Países      GDP    PCs
1      Afghanistan 1261.35 4.58
2           Algeria 6354.64 14.00
3          Andorra 31630.74 81.00
4 Antigua and Barbuda 13508.80 82.00
5          Argentina 15595.39 47.70
6          Australia 35253.94 78.95
```

En cuanto al resumen de la variable PCs, tendremos en cuenta que se trata de una variable cuantitativa continua. Dado el tipo de variable, tal como vimos en la práctica anterior, los **gráficos** adecuados para la misma han de ser el histograma y el diagrama de caja:

```
> attach(datos)
> boxplot(PCs,main="Resumen gráfico de la variable PCs",col="yellow")
> hist(PCs,main="Resumen gráfico de la variable PCs",col="purple")
```

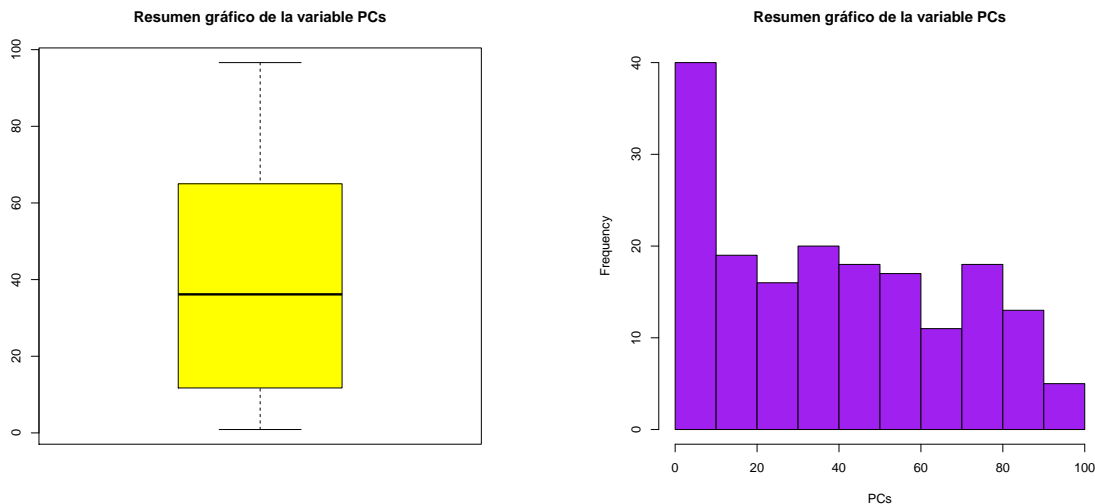


Figura 1: Diagrama de caja (izquierda) e histograma (derecha) de la muestra de la variable PCs.

Para estudiar la relación existente entre PCs y GDP calcularemos la covarianza y la correlación. Hay que tener presente que la función `cov` calcula la cuasicovarianza, es decir, con el factor $\frac{1}{n-1}$, por ello es necesario "corregirlo":

```
> ((n-1)/n)*cov(PCs,GDP)
```

```
[1] 362855.2
```

Esto no ocurre con la correlación, que la podemos calcular directamente con `cor`:

```
> cor(PCs,GDP)
```

```
[1] 0.7738078
```

El valor positivo de ambas cantidades indica una relación positiva (creciente) entre las mismas, además un valor relativamente alto de la correlación (0.77) indica que esa relación es “bastante lineal”. Prueba a dibujar simultáneamente ambas variables en un gráfico de dispersión para comprobar visualmente la existencia de esa relación lineal.

Ahora realizamos el cálculo de la **varianza y la desviación típica** teniendo en cuenta que las funciones `var` y `sd` proporcionan la cuasivarianza y cuasidesviación típica respectivamente:

```
> n=length(PCs)
> n
[1] 177
> varianza=var(PCs)*(n-1)/n
> varianza
[1] 812.5306
> desvtip<-sqrt(varianza)
> desvtip
[1] 28.50492
```

El siguiente paso es el cálculo de los **cuantiles**:

```
> prob=c(0.3,0.6,0.9)
> quantile(PCs,prob)
 30%   60%   90%
16.780 46.314 80.250
```

El proceso de **tipificación** consiste en restar la media y dividir por la desviación típica. Crea una variable que no tiene unidades, por tanto, es útil para eliminar la influencia de las mismas en el análisis. Para tipicar la variable procederíamos como sigue

```
> media_PCs<-mean(PCs)
> media_PCs
[1] 39.16367
> var_PCs<-var(PCs)*(length(PCs)-1)/length(PCs)
> var_PCs
[1] 812.5306
> PCs_tipi<-(PCs-media_PCs)/sqrt(var_PCs)
```

En el último apartado se pide repetir lo mismo, pero con la variable GDP que es cuantitativa continua, por lo que el código sería análogo.

2. La recta de regresión

Para modelizar y explicar la dependencia de una variable respuesta Y con respecto a una variable explicativa X , utilizaremos los modelos de regresión. En esta práctica, consideraremos el **modelo de regresión lineal simple**. En general, los modelos de regresión se diseñan con dos objetivos: primero, conocer de qué modo Y depende de X y, una vez establecido el modelo de dependencia, utilizarlo para predecir valores de Y a partir de observaciones de X .

En la Figura 2, podemos ver distintos tipos de relaciones entre las variables X e Y (representamos los puntos $\{(X_i, Y_i)\}_{i=1}^n$). En la primera gráfica, vemos una relación lineal entre la variable explicativa y la variable respuesta, de manera que a mayores valores de X le corresponden mayores valores de Y . En la segunda gráfica, la relación, para valores pequeños de X es parabólica, pero no lineal. En la última gráfica, podemos ver que a medida que aumenta el valor de la X , la dispersión de las observaciones con respecto a un posible ajuste lineal se incrementa.

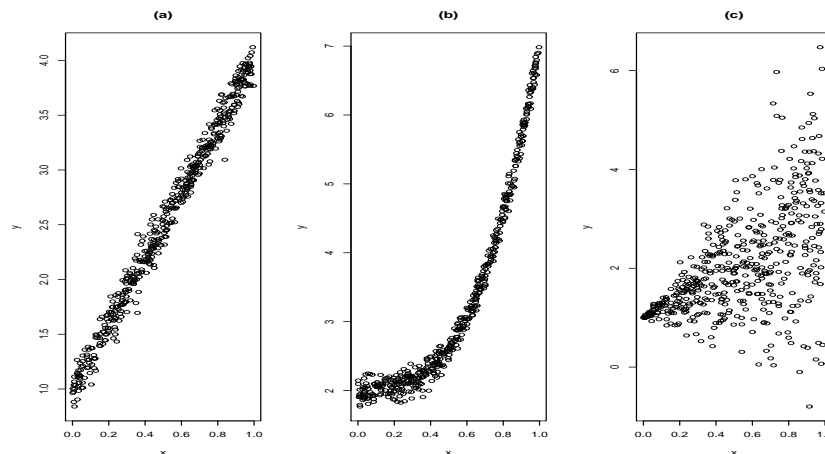


Figura 2: Diagramas de dispersión para distintos tipos de relaciones entre variables.

2.1. El modelo lineal

El modelo de regresión lineal relaciona una variable explicativa X con una variable respuesta Y a través de la siguiente expresión:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \in N(0, \sigma_\varepsilon^2)$$

donde β_0 es la ordenada en el origen o intercepto, β_1 es la pendiente (indica la razón de cambio en Y cuando X varía en una unidad) y ε es el término de error, que sigue una distribución normal de media 0 y varianza constante σ_ε^2 (modelo homocedástico). Aquellos modelos donde la varianza no se mantiene constante (tercera gráfica de la Figura 2), se denominan heterocedásticos.

2.2. Estimación por mínimos cuadrados

En la práctica, dispondremos de una muestra $\{(x_1, y_1), \dots, (x_n, y_n)\}$ de valores de (X, Y) y determinaremos una estimación de los parámetros del modelo β_0 y β_1 . El método que utilizaremos para obtener nuestros estimadores es el **Método de Mínimos Cuadrados**.

Cuando tenemos unos estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ los errores de predicción o **residuos** (denotados por e_i) proporcionan la diferencia entre los valores observados de la variable Y y los que ajusta el modelo:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \dots, n.$$

Los estimadores de mínimos cuadrados de β_0 y β_1 son aquellos valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la suma de cuadrados de los residuos:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

En la Figura 3 los segmentos verticales representan la diferencia entre el valor observado y el valor que predeciría una recta ajustada (lo que hemos denominado residuos). El objetivo es seleccionar la recta que minimice la suma de estos residuos al cuadrado.

Además, estos estimadores tienen la siguiente expresión explícita:

$$\hat{\beta}_1 = \frac{S_{X,Y}}{s_X^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

donde \bar{y} y \bar{x} denotan las medias muestrales de y_1, \dots, y_n y x_1, \dots, x_n , respectivamente; s_X^2 es la varianza muestral de X :

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

y $S_{X,Y}$ es la **covarianza** muestral entre X e Y , que viene dada por:

$$S_{X,Y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

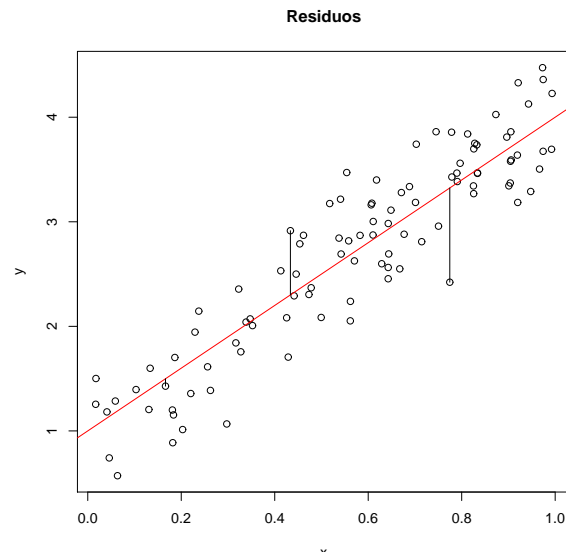


Figura 3: Residuos a minimizar en el método de mínimos cuadrados. Los segmentos verticales representan los residuos e_i .

En la Figura 4, en el gráfico de la izquierda, representamos la línea ajustada, con $\hat{\beta}_0$ y $\hat{\beta}_1$ obtenidos por el método de mínimos cuadrados.

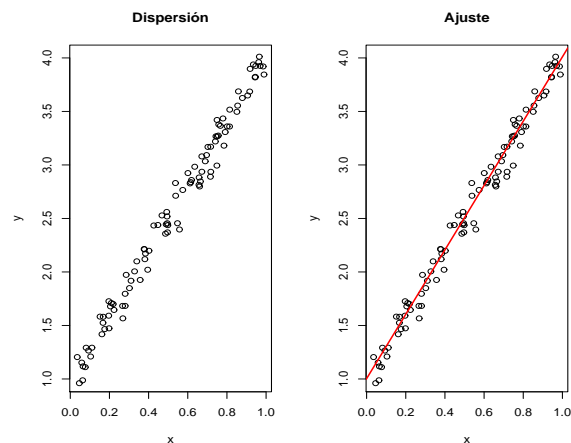


Figura 4: Diagrama de dispersión y ajuste lineal.

En el modelo lineal, una vez determinados los estimadores de la pendiente $\hat{\beta}_1$ y el intercepto $\hat{\beta}_0$ de

la recta, nos quedaría por determinar la varianza del error. La podemos obtener como:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2.$$

La función `lm` permite estimar un modelo de regresión lineal con . Si almacenamos el resultado de la función en una variable denominada `modelo`, podemos ver que devuelve, entre otras variables: los coeficientes del modelo (`coefficients`), los residuos (`residuals`) y los valores ajustados (`fitted.values`).

2.3. Algunos coeficientes de interés

Se denomina **coeficiente de regresión** a la pendiente (parámetro $\hat{\beta}_1$) de la recta de regresión de Y sobre X . Este coeficiente proporciona información sobre el comportamiento de la variable respuesta Y en función de la variable explicativa X . Así tenemos:

- a) Si $\hat{\beta}_1$ es cero (o próximo a cero), para cualquier valor de X la variable Y no presenta una tendencia lineal como función de X , distribuyéndose como un error aleatorio alrededor del valor β_0 .
- b) Si $\hat{\beta}_1 > 0$, al aumentar los valores de X también tenderán a aumentar los valores de Y . Es decir, existe una relación directa entre las variables.
- c) Si $\hat{\beta}_1 < 0$, al aumentar X , los valores de Y disminuyen. Es decir, existe una relación inversa entre las variables.

El **coeficiente de correlación lineal** entre X e Y cuantifica la dependencia lineal que existe entre las dos variables y se define como:

$$r = \frac{S_{X,Y}}{s_X s_Y}.$$

En , podríamos calcularlo como:¹

```
cor(x,y)
cov(x,y)/(sd(x)*sd(y))
```

Este coeficiente no tiene dimensiones y toma valores en $[-1, 1]$. Valores cercanos a 1 nos indicarían una relación lineal directa, mientras que valores cercanos a -1 darían una relación lineal inversa.

¹Recordemos que en las funciones `sd` y `cov` incluyen el coeficiente $\frac{1}{n-1}$ en lugar de $\frac{1}{n}$.

Si las variables son independientes, entonces $r \approx 0$. Sin embargo, el recíproco no es cierto, ya que este coeficiente sólo cuantifica relaciones lineales.


Finalmente, el cuadrado del coeficiente de correlación lineal, r^2 se denomina **coeficiente de determinación** y mide la proporción de variabilidad de Y que explica X , proporcionándonos una medida del ajuste del modelo. Si r^2 toma valores próximos a 1, esto será indicativo de un buen ajuste. El coeficiente de determinación del modelo de regresión lineal simple viene dado por:

$$r^2 = \frac{S_{X,Y}^2}{s_X^2 s_Y^2}.$$

Ejercicio. Estudia la dependencia que presenta la variable GDP en función de la variable PCs. Es decir, ¿podríamos explicar el comportamiento de la variable GDP en función de la variable PCs?

1. Representa un diagrama de dispersión de GDP sobre PCs.
2. Estudia la relación entre GDP y PCs a través de la covarianza y del coeficiente de correlación.
3. Ajusta un modelo de regresión lineal de la variable GDP sobre PCs. Representa el modelo ajustado sobre el diagrama de dispersión obtenido en el apartado a).
4. Obtén el coeficiente de determinación e interpreta el valor obtenido.
5. Representa los residuos del modelo frente a la variable explicativa para testar la hipótesis de homocedasticidad.
6. Calcula predicciones para la variable GDP sabiendo que la variable PCs toma los valores $x_0 = 5$, $x_0 = 25$ y $x_0 = 75$.
7. Ajusta y representa de nuevo el modelo de regresión de GDP sobre PCs, pero únicamente teniendo en cuenta los datos de países de la Unión Europea. Para facilitar la tarea de selección, se proporciona a continuación el listado de los mismos:

```
UE=c("Austria", "Belgium", "Bulgaria", "Croatia", "Republic of Cyprus",  
     "Czech Republic", "Denmark", "Estonia", "Finland", "France", "Germany",  
     "Greece", "Hungary", "Ireland", "Italy", "Latvia", "Lithuania",  
     "Luxembourg", "Malta", "Netherlands", "Poland", "Portugal", "Romania",  
     "Slovakia", "Slovenia", "Spain", "Sweden")
```

Vamos a realizar con  las distintas cuestiones que se piden. Comenzaremos con la representación del diagrama de dispersión:

```
> plot(PCs,GDP,main="Diagrama de dispersión",pch=19)
```

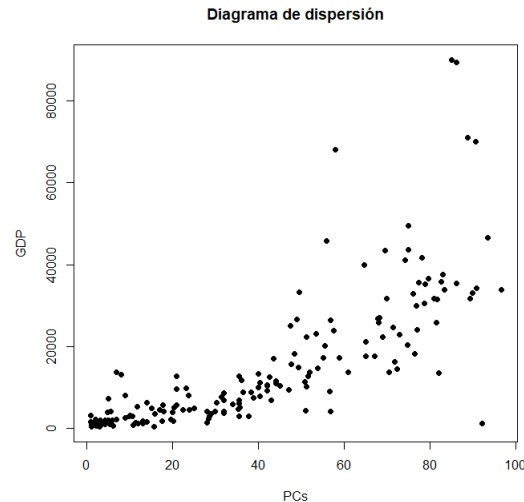


Figura 5: Diagrama de dispersión de las variables PCs y GDP.

El ajuste del modelo, se hará con la función `lm`, añadiremos la recta al diagrama de dispersión y obtendremos un resumen del modelo, en el cual se incluye el coeficiente de determinación

```
> modelo<-lm(GDP~PCs)
```

```
> modelo
```

Call:

```
lm(formula = GDP ~ PCs)
```

Coefficients:

```
(Intercept)      PCs  
-3032.6       446.6
```

```
> plot(PCs,GDP,main="Diagrama de dispersión",pch=19)
```

```
> abline(beta0,beta1,col=2,lwd=2)
```

```
> legend(x="topleft",legend="Ajuste lineal",lwd=2,col=2)
```

```
> summary(modelo)
```

Call:

```
lm(formula = GDP ~ PCs)
```

Residuals:

Min	1Q	Median	3Q	Max
-36881	-5484	-824	3030	55113

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3032.62	1338.53	-2.266	0.0247 *
PCs	446.57	27.63	16.161	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10480 on 175 degrees of freedom

Multiple R-squared: 0.5988, Adjusted R-squared: 0.5965

F-statistic: 261.2 on 1 and 175 DF, p-value: < 2.2e-16

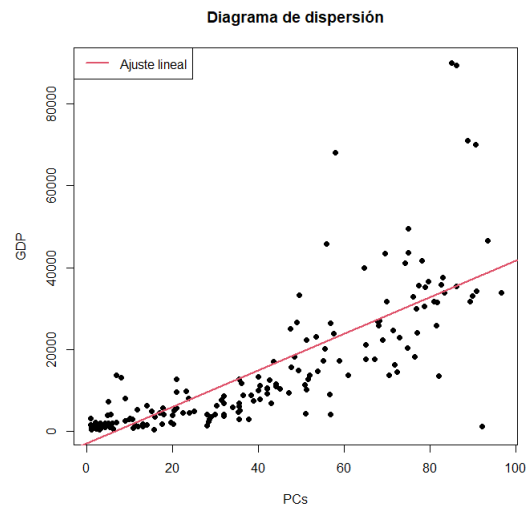


Figura 6: Diagrama de dispersión junto con el modelo de regresión lineal simple ajustado.

A continuación procederemos al cálculo y representación de los residuos:

```
> residuos<-modelo$residual
> plot(PCs,residuos,main="Representación de los residuos",pch=19)
```

```
> abline(h=0,col="gray",lwd=2)
```

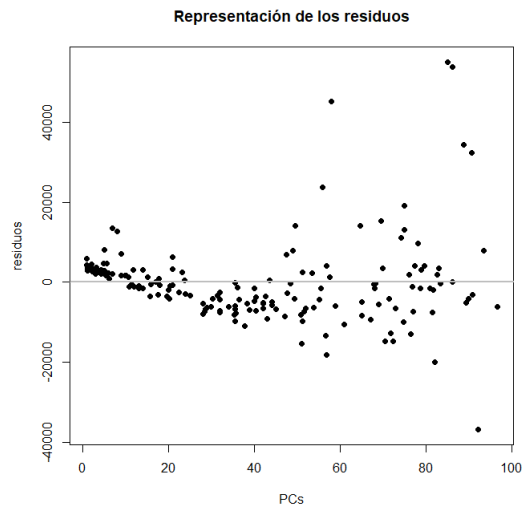



Figura 7: Diagrama de los residuos resultantes del ajuste del modelo de regresión lineal simple.

Para la realización de las predicciones tenemos que tener en cuenta que los valores en los que vamos a predecir la variable respuesta tienen que hallarse dentro del rango de los valores de la variable explicativa.

La primera opción para realizar este cálculo no es más que sustituir los valores en la ecuación de la recta de regresión.

```
> x0<-c(5,25,75)
> pred<-beta0+beta1*x0
> pred
[1] -799.7502 8131.7330 30460.4411
```

También podemos hacerlo usando la función `predict` de  que nos aporta algo más de información.

```
> x0 <- data.frame(PCs =c(5,25,75))
> predict(modelo,x0)
      1      2      3
-799.7502 8131.7330 30460.4411
```

Para la realización del último apartado, necesitamos hacer una selección dentro de la base de datos:

```
> UE=c("Austria", "Belgium", "Bulgaria", "Croatia", "Republic of Cyprus",  
+ "Czech Republic", "Denmark", "Estonia", "Finland", "France", "Germany",  
+ "Greece", "Hungary", "Ireland", "Italy", "Latvia", "Lithuania",  
+ "Luxembourg", "Malta", "Netherlands", "Poland", "Portugal", "Romania",  
+ "Slovakia", "Slovenia", "Spain", "Sweden")  
> pp=datos$Países[datos$Países %in% UE]  
> pos=which(datos$Países %in% UE)  
> PCs_UE=PCs[pos]  
> GDP_UE=GDP[pos]
```

Una vez hecho esto, el código es análogo al de los apartados anteriores, pero empleando estas nuevas variables que acabamos de definir.