ESMA 4016
PREDICCION: REGRESIÓN LINEAL

Dr. Edgar Acuña
http://academic.uprm.edu/eacuna

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ

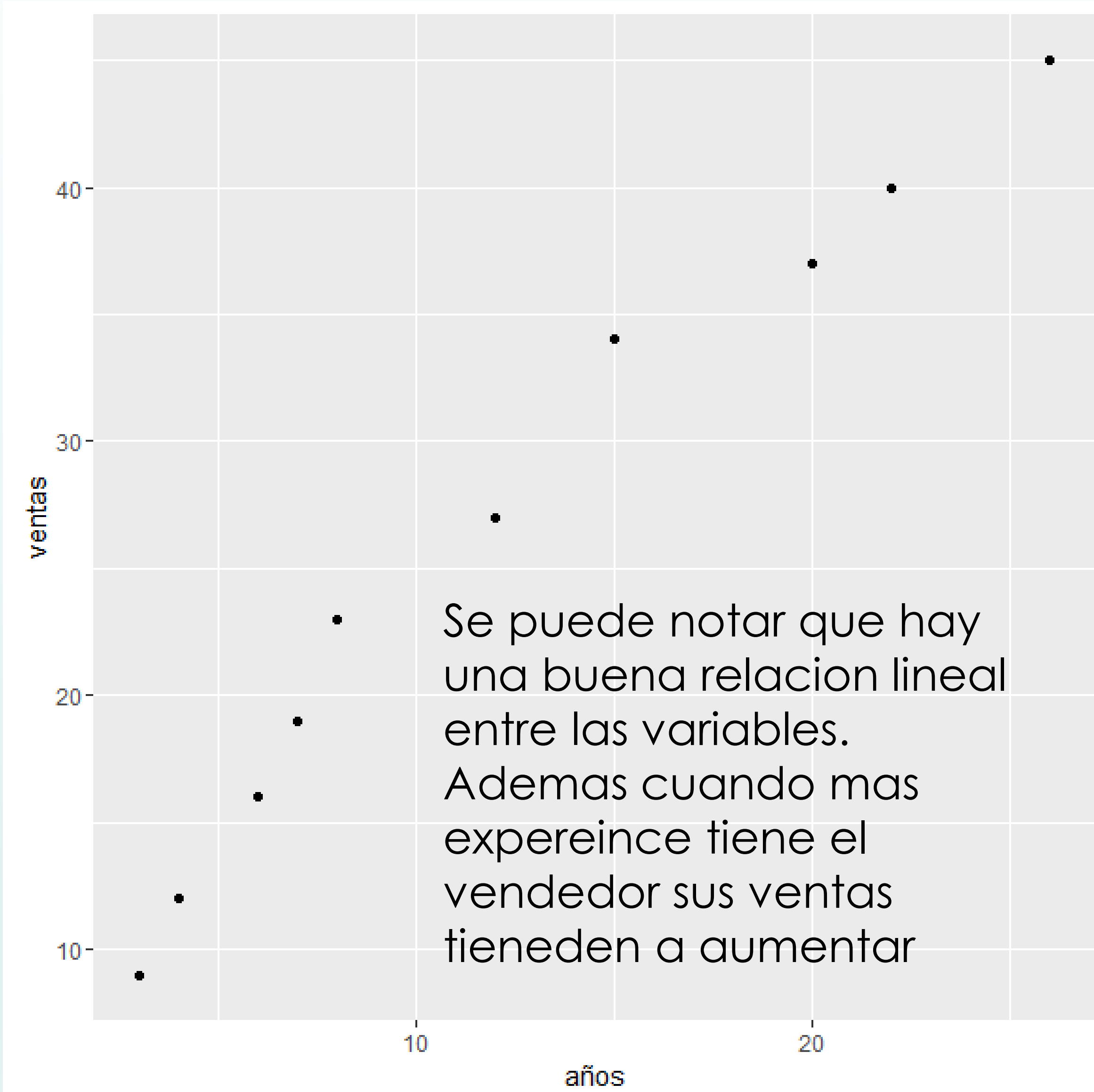
Ejemplo 3.1

Ejemplo 3.1. El dueño de una empresa que vende carros desea determinar si hay relación lineal entre los años de experiencia de sus vendedores y la cantidad de carros que venden. Los siguientes datos representan los años de experiencia (X) y las unidades de carros vendidas al año (Y), de 10 vendedores de la empresa.

X(años)	3	4	6	7	8	12	15	20	22	26
Y(ventas)	9	12	16	19	23	27	34	37	40	45

Solución:

Primero hacemos un plot considerando los años de experiencia en el eje horizontal y las ventas en el eje vertical. En **Python se usa la funcion scatter** . En la librería plotline se usa ggplot junto con la opcion geom_point().



Se puede notar que hay una buena relacion lineal entre las variables. Ademas cuando mas expereince tiene el vendedor sus ventas tieneden a aumentar

3.2 El Coeficiente de Correlación

Llamado también coeficiente de correlación de Pearson, se representa por **r** y es una medida que representa el grado de asociación entre dos variables cuantitativas X e Y.

Se calcula por

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Donde:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \quad , \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \quad \text{y} \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Tanto S_{xx} como S_{yy} no pueden ser negativas, S_{xy} si puede ser positiva o negativa.

- La correlacion varia entre -1 y 1
- En la mayoria de los problemas, una correlacion mayor que .75 o menor que -.75 es considerada bastante aceptable. Una correlacion que cae entre -.3 y .3 es considerada muy baja.
- Si la correlacion es positiva entonces cuando X aumenta se espera que Y tambien aumente.
- Si la correlacion es negativa entonces cuando X aumenta se espera que Y disminuya.

Ejemplo (cont)

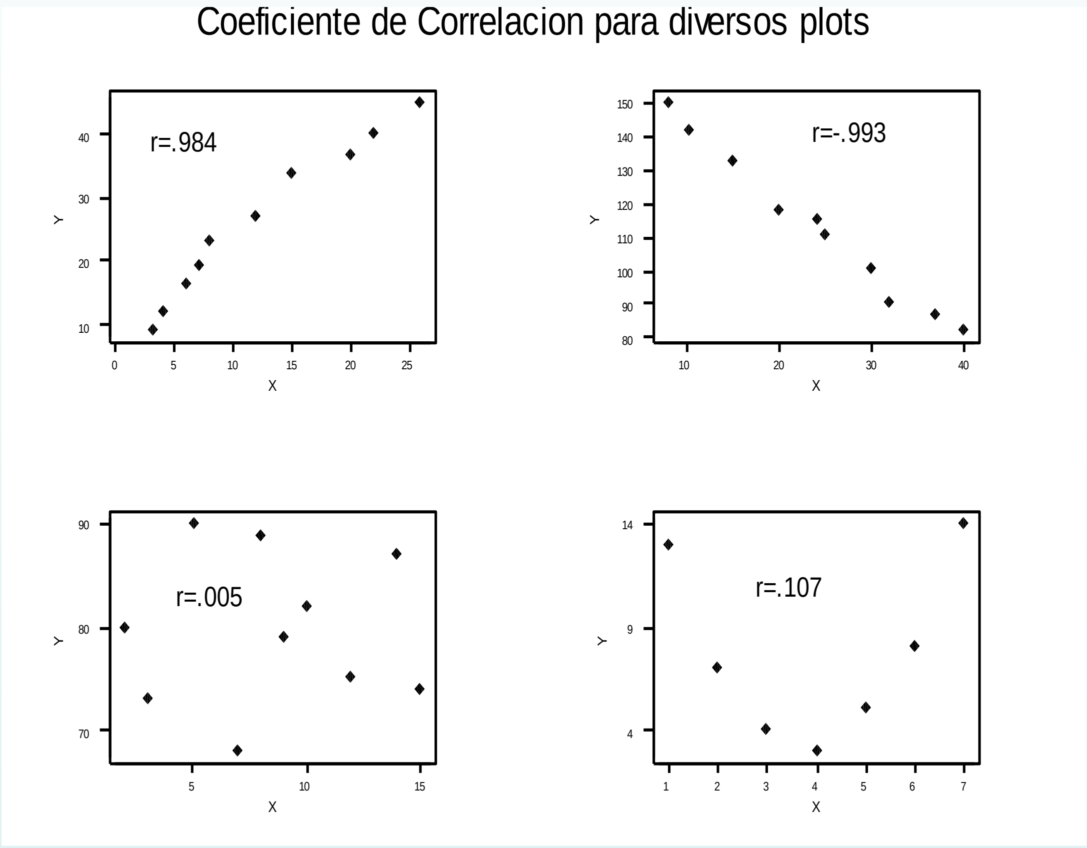
Row	years	ventas	Sxx	Syy	Sxy	r
1	3	9	590.1	1385.6	889.4	0.983593
2	4	12				
3	6	16				
4	7	19				
5	8	23				
6	12	27				
7	15	34				
8	20	37				
9	22	40				
10	26	45				

En **numpy, pandas y stats**, el coeficiente de correlación se puede obtener usando la funcion corrcoef y corr**.**

df.corr()["years"]["ventas"]

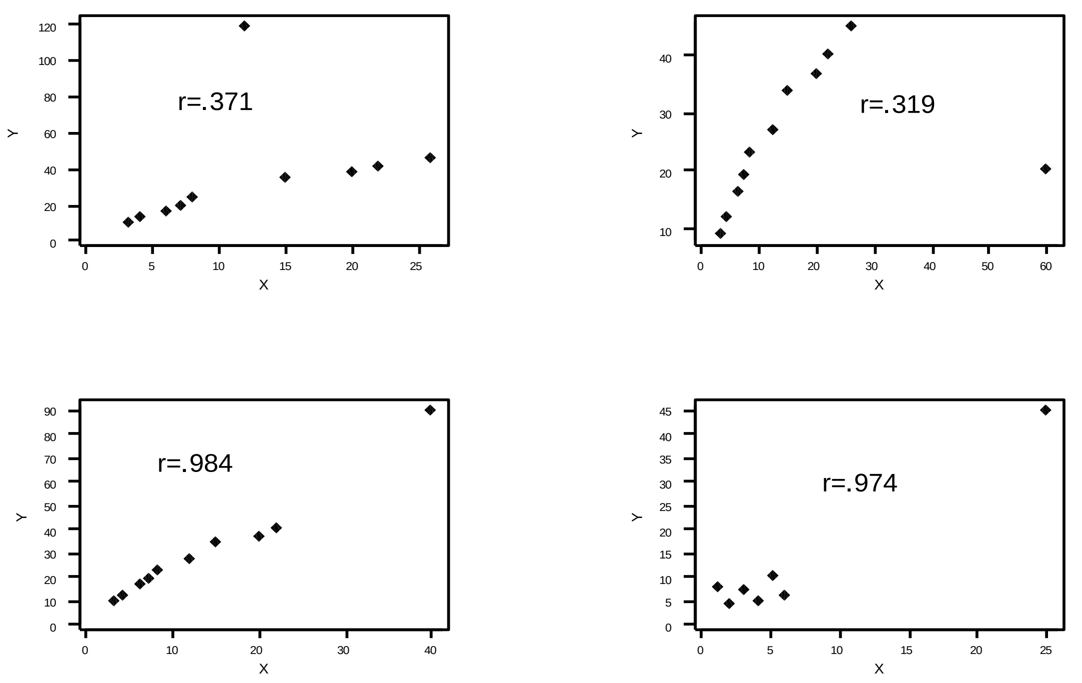
0.9835928893659418

Interpretación: *Existe una buena relación lineal entre los años de experiencia y las unidades que vende el vendedor. Además mientras más experiencia tiene el vendedor más carros venderá. Se puede usar los años de experiencia para predecir las unidades que venderá anualmente a través de una línea recta.*



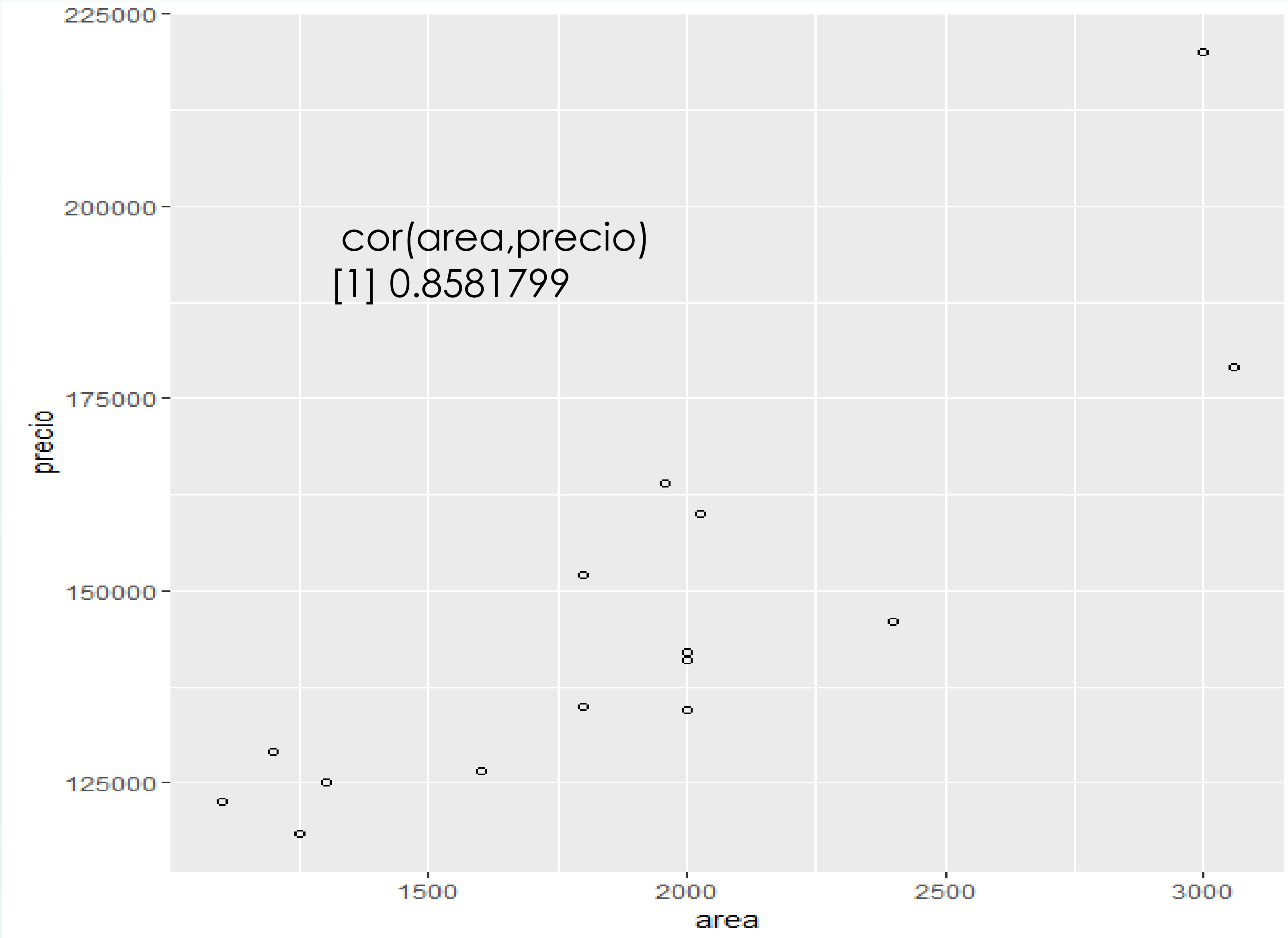
Coeficiente de Correlacion para diversos plots
r=.984
r=-.993
r=.005
r=.107
X
Y

Efecto de valores anormales en el valor de la correlacion



Ejemplo 3.2

Casa	**área(pies^2)**	**precio**
1	3060	179000
2	1600	126500
3	2000	134500
4	1300	125000
5	2000	142000
6	1956	164000
7	2400	146000
8	1200	129000
9	1800	135000
10	1248	118500
11	2025	160000
12	1800	152000
13	1100	122500
14	3000	220000
15	2000	141000


cor(area,precio)
[1] 0.8581799
225000
200000
175000
150000
125000
precio
1500
2000
2500
3000
area

Regresión Lineal Simple

Se trata de predecir el comportamiento de Y usando X entonces el **modelo de regresión lineal simple** es de la forma:

$$Y = \alpha + \beta X + \varepsilon$$

Donde, Y es llamada la variable de respuesta o dependiente,

X es llamada la variable predictora o independiente,

α es el intercepto de la línea con el eje Y,

β es la pendiente de la línea de regresión y

ε es un error aleatorio, el cual se supone que tiene media 0 y varianza constante σ^2.

Línea de regresión estimada

El modelo de regresion lineal es estimado por la ecuacion

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

El estimado $\hat{\alpha}$ de α y el estimado $\hat{\beta}$ de β son hallados usando el método de mínimos cuadrados, que se basa en minimizar la suma de cuadrados de los errores.

$$Q(\alpha,\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Luego se obtienen $\hat{\beta} = \dfrac{s_{xy}}{s_{xx}}$ y $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Resultados del Ejemplo 3.1

```
Intercept 7.661413
years 1.507202
dtype: float64
OLS Regression
    ==========================================================
    Dep. Variable: ventas                R-squared: 0.967
    Model: OLS                           Adj. R-squared: 0.963
    Method: Least Squares                F-statistic: 237.8
    Date: Thu, 08 Mar 2018               Prob (F-statistic): 3.11e-07
    Time: 13:09:36                       Log-Likelihood: -21.720
    No. Observations: 10                 AIC: 47.44
    Df Residuals: 8                      BIC: 48.05
    Df Model: 1
    Covariance Type: nonrobust
    ==========================================================
    coef                std err    t      P>|t|      [0.025 0.975]
    Intercept 7.6614    1.417    5.405   0.001       4.393 10.930
    years     1.5072    0.098   15.421   0.000       1.282 1.733
```

Resultados del Ejemplo 3.2

```
Intercept 73167.748381
area 38.523071
dtype: float64
OLS Regression Results
==============================================================
Dep. Variable: precio                  R-squared: 0.736
Model: OLS                             Adj. R-squared: 0.716
Method: Least Squares                  F-statistic: 36.33
Date: Thu, 08 Mar 2018                 Prob (F-statistic): 4.25e-05
Time: 13:09:37                         Log-Likelihood: -163.54
No. Observations: 15                   AIC: 331.1
Df Residuals: 13                       BIC: 332.5
Df Model: 1
Covariance Type: nonrobust
======================================================================
coef                 std err     t      P>|t|    [0.025 0.975] –
Intercept  7.317e+04 1.27e+04 5.773  0.000   4.58e+04 1.01e+05
area       38.5231      6.391  6.028  0.000     24.716 52.330
```

Interpretación de los Coeficientes de Regresión:

- **Interpretación del intercepto $\hat{\alpha}$:**

Indica el valor promedio de la variable de respuesta Y cuando X es cero. Si se tiene certeza de que la variable predictora X no puede asumir el valor 0,

entonces la interpretación no tiene sentido.

En el ejemplo anterior, $\hat{\alpha}$ = 73,168 indicaría que si la casa no tiene área, su precio promedio será 73,158, lo cual no es muy razonable.

- **Interpretación de la pendiente $\hat{\beta}$:**

Indica el cambio promedio en la variable de respuesta Y cuando X se incrementa en una unidad.

En el ejemplo anterior $\hat{\beta}$ = 38.5 indica que por cada pie cuadrado adicional de la casa su precio aumentará en promedio en 38.5 dólares.

Inferencia en Regresión Lineal

- **Inferencia acerca de los coeficientes de regresión**

Las pruebas de hipótesis más frecuentes son, Ho: $\alpha = 0$ versus Ha: $\alpha \neq 0$ y

Ho: $\beta = 0$ versus Ha: $\beta \neq 0$.

La prueba estadística para el caso de la pendiente viene dada por:

$$t = \frac{\hat{\beta}}{s.e(\hat{\beta})} = \frac{\hat{\beta}}{\frac{s}{\sqrt{S_{xx}}}} \quad \text{y} \quad s = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}}$$

La cual se distribuye como una *t* con *n*-2 grados de libertad.

En **R** aparece el valor de la prueba estadística y el “p-value” de la prueba, el cual se puede usar para llegar a una decisión. Un "p-value" cercano a 0, digamos menor que 0.05, lleva a la conclusión de rechazar la hipótesis nula.

Si se rechaza la hipótesis nula quiere decir de que de alguna manera la variable X es importante para predecir el valor de Y usando la regresión lineal. En cambio si se acepta la hipótesis nula se llega a la conclusión de que, la variable X no es importante para predecir el comportamiento de Y usando una regresión lineal.

En el Ejemplo 3.2 el valor de la prueba estadística de *t* es 6.028 y el P-value = .0000 por lo que se rechaza la hipótesis nula. Luego hay suficiente evidencia estadística para concluir que la variable área de la casa puede ser usada para predecir el precio de la casa.

El Coeficiente de Determinación

Es una medida de la bondad de ajuste del modelo de regresión hallado.

$$R^2 = \frac{SSR}{SST}$$

Donde,

SSR representa la suma de cuadrados debido a la regresión y

SST representa la suma de cuadrados del total.

El coeficiente de determinación es simplemente el cuadrado del coeficiente de correlación. El coeficiente de Determinación varía entre 0 y 1, aunque es bastante común expresarlo en porcentaje. Un R^2 mayor del 70 % indica una buena asociación lineal entre las variables, luego la variable X puede usarse para predecir Y. ***R^2*** indica qué porcentaje de la variabilidad de la variable de respuesta Y es explicada por su relación lineal con X.

En el ejemplo salio R^2=73.6 esto significa que solo el 73.6% de la variabilidad de los precios de las casas es explicada por su relacion lineal con el area de la misma. Se podria usar el area de la casa para predecir su precio.

Regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras $X_1,\ldots X_p$, es de la siguiente forma:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_pX_p + \varepsilon$$

Las constantes $b_0, b_1, \ldots, b_p$, llamadas coeficientes de regresión, se estiman usando el método de mínimos cuadrados, y usando *n* observaciones de la forma $y_i, x_{i1}, x_{i2}, \ldots, x_{ip}$, donde $i = 1, \ldots, n$. La cantidad ε es una variable aleatoria con media 0 y varianza σ^2.

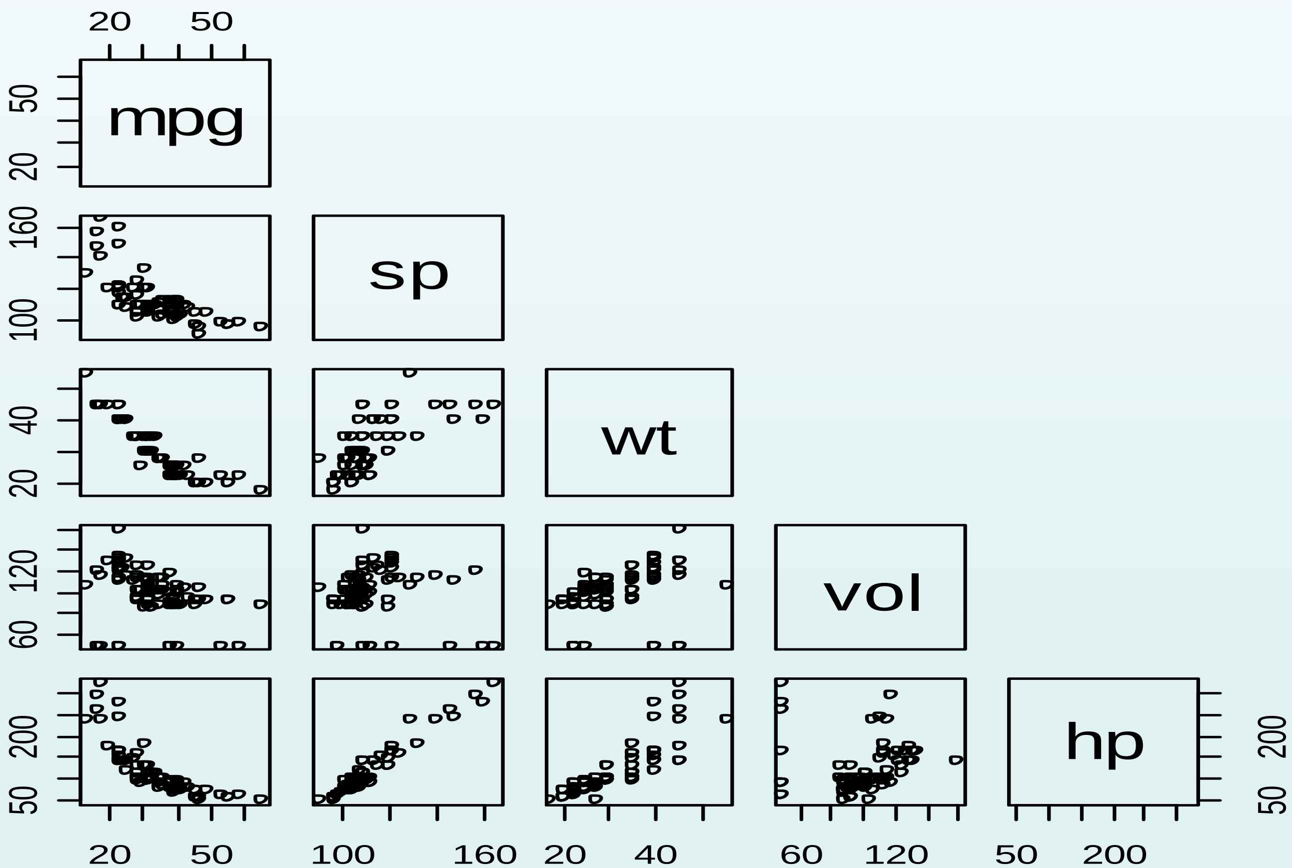
Interpretación del coeficiente de regresión estimado β_j

El estimado del coeficiente de regresión poblacional b_j, con $j = 1, \ldots, p$, se representará por β_j. Este estimado indica el cambio promedio en la variable de respuesta Y cuando la variable predictora X_j cambia en una unidad adicional asumiendo que las otras variables predictoras permanecen constantes.

Ejemplo

Considerar el conjunto de datos **millaje** donde la variable de respuesta es
Y= (MPG) millas promedio por galón de un auto, y las variables predictoras son;
X_1=(VOL): Capacidad en volumen del carro,
X_2=(HP): Potencia del Motor,
X_3= (SP) :Velocidad Máxima y
X_4=(WT): Peso del auto.
Los datos están dsponibles en academic.uprm.edu/eacuna/millaje.txt.

Plot matricial



Edgar Acuña

El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones está dado por:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + e_i \quad \text{para i = 1,2,…,n}$$

en forma matricial : $\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ .. & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Estimación del vector de parámetros β por Cuadrados Mínimos

Se tiene que minimizar la suma de cuadrados de los errores.

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Derivando Q con respecto a **β** e igualando a cero se obtiene el sistema de ecuaciones normales

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

resolviendo para **β** se obtiene el vector de coeficientes estimados

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Estimación del vector de parámetros β (2)

En R se usa la funcion lm(formula, data),

Donde formula es de la forma y~x1+x2+…..+xp , y es la variable de respuesta y x1, x2,…xp son las variables predictoras.

En Python se usa el modulo statmodels, que tiene la function ols la cual halla los estimados de los coeficientes de regression.

import statsmodels.formula.api as smf

smf.ols(formula= ‘y~x1+x2+…..+xp’,data).fit()

Estimación del vector de parámetros β (2)

- El modelo de regresión que se obtiene para el ejemplo es el siguiente:

  MPG = 192 - 1.29 SP - 1.85 WT -0.01 VOL + 0.39 HP
- Interpretacion de los coeficientes:
- B1=-1.29 significa que si la velocidad maxima aumenta en una unidad(m/h)se espera que el rendimiento en millas por galon baje en 1.29 asumiendo que las otras variables permanecen constante.
- B2=-1.85 significa que si el peso del carro aumenta en una unidad (100 lb)se espera que el rendimiento en millas por galon disminuya en 1.85 asumiendo que las otras variables permanecen constante.
- B3=-.01 significa que si el volumen interior del carro aumenta en una unidad (pie cubico) se espera que el rendimiento en millas por galon disminuya en .01 asumiendo que las otras variables permanecen constante.
- B4=0.39 significa que si la potencia del motor aumenta en una unidad se espera que que el rendimiento en millas por galon aumente en .39 asumiendo que las otras variables permanecen constante.

Inferencia en Regresión lineal múltiple

Involucra realizar

- Pruebas de hipótesis e intervalos de confianza acerca de los coeficientes del modelo de regresión poblacional.
- Intervalos de confianza de las predicciones que se hacen con el modelo.

2.3.1 Prueba de hipótesis acerca de un coeficiente de regresión individual

- Ho: $\beta_i = 0$ (i=1,2,..,p),
- Ha: $\beta_i \neq 0$;

La prueba estadística es la prueba de t:

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{C_{ii}}}$$

se distribuye como una t con (n-p-1) gl.

Donde, C_{ii} es el i-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$.

Los programas de computadoras, dan el “P-value” de la prueba t. Si el P-value es <0.05 se rechaza Ho

Selección de Variables en Regresion Lineal Multiple

- Método de eliminación hacia atrás (“Backward Elimination”)
- Método de Selección hacia adelante (“Forward Selection”):
- Método Paso a Paso ("Stepwise")

Método de eliminación hacia atrás

Aquí en el paso inicial se incluyen en el modelo a todas las variables predictoras y en cada paso se elimina la variable menos importante de acuerdo a varios criterios, Por ejmplo

La variable cuyo "p-value" es más grande para la prueba de *t* o cuyo valor de la prueba *t* menor que 2 en valor absoluto.

Una variable que es eliminada del modelo ya no puede volver a entrar en un paso subsiguiente.

El proceso termina cuando todos los "p-values" son menores que .05, o cuando todos los valores de la prueba t son mayores que 2 en valor absoluto.

R/Python usan en lugar de la prueba de t el criterio de información de Akaike (AIC)

El criterio AIC

El criterio de información de Akaike (Akaike, 1973), tiene su origen en conceptos de teoría de información y está basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta *Y* bajo el modelo reducido y bajo el modelo completo. Se define como,

$$AIC= n*\log[SSE_p/n]+2p \qquad (1)$$

Donde n, es el numero de casos, p es el número de parámetros del modelo, y SSE es la suma de cuadros de los errores de la regresión Existen otras variantes a la formula (1). Statmodels usa

$$AIC= n*[2*\pi*\log(SSE_p/n-p)+1]+p$$

Un buen modelo es aquel con bajo AIC.

Ejemplo 3.6. El conjunto de datos **grasa** contiene 13 variables que sirven para predecir el porcentaje de grasa en el cuerpo humano

Columna	Nombre
v1	grasa (% de grasa)
v2	edad (en años)
v3	peso (en libras)
v4	altura (en pulgadas)
v5	cuello (en cms)
v6	pecho (en cms)
v7	abdomen (en cms)
v8	cadera (en cms)
v9	muslo (en cms)
v10	rodilla (en cms)
v11	tobillo (en cms)
v12	biceps (en cms)
v13	antebrazo (en cms)
v14	muñeca (en cms)

Se tomaron las mediciones en 252 sujetos.

```
> #Hallando el mejor subcojunto usando stepwise (backward) y el criterio AIC
> l1<-lm(grasa~.,data=grasa)
> step(l1,scope=~.,direction="backward")
Start:  AIC=749.36
grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + rodilla + tobillo + biceps + antebrazo + muneca
             Df Sum of Sq    RSS    AIC
- rodilla     1      0.07   4411.5    747.36
- pecho       1      1.07   4412.5    747.42
- altura      1      9.74   4421.2    747.91
- tobillo     1     11.44   4422.9    748.01
- biceps      1     20.87   4432.3    748.55
<none>                      4411.4    749.36
- cadera      1     37.50   4448.9    749.49
- muslo       1     49.58   4461.0    750.17
- peso        1     50.61   4462.1    750.23
- edad        1     68.26   4479.7    751.23
- cuello      1     75.96   4487.4    751.66
- antebrazo   1     95.51   4507.0    752.76
- muneca      1    170.12   4581.6    756.89
- abdomen     1   2260.95   6672.4    851.63
```

La variable rodilla es la primera variable en ser eliminada del modelo por tener el AIC mas bajo

Step: AIC=747.36
grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + tobillo + biceps + antebrazo + muneca

	Df	Sum of Sq	RSS	AIC
- pecho	1	1.13	4412.7	745.43
- altura	1	9.66	4421.2	745.91
- tobillo	1	12.09	4423.6	746.05
- biceps	1	20.81	4432.3	746.55
<none>			4411.5	747.36
- cadera	1	37.43	4448.9	747.49
- peso	1	53.08	4464.6	748.38
- muslo	1	54.88	4466.4	748.48
- edad	1	74.06	4485.6	749.56
- cuello	1	78.44	4490.0	749.80
- antebrazo	1	96.77	4508.3	750.83
- muneca	1	170.55	4582.1	754.92
- abdomen	1	2269.88	6681.4	849.97

- La variable Pecho es la segunda variable en ser eliminada del modelo

Y asi se siguen eliminado variables una por uno hasta llegar al modelo final
grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo + muneca

	Df	Sum of Sq	RSS	AIC
<none>			4455.3	741.85
- cadera	1	36.5	4491.8	741.91
- cuello	1	79.1	4534.4	744.29
- edad	1	83.8	4539.1	744.55
- peso	1	93.0	4548.3	745.05
- muslo	1	100.7	4556.0	745.48
- antebrazo	1	140.5	4595.8	747.67
- muneca	1	166.8	4622.2	749.12
- abdomen	1	3163.0	7618.3	875.04

Call:
lm(formula = grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo + muneca, data = grasa)

Coefficients:

(Intercept)	edad	peso	cuello	abdomen	cadera
-22.65637	0.06578	-0.08985	-0.46656	0.94482	-0.19543

muslo	antebrazo	muneca
0.30239	0.51572	-1.53665

Interpretación: El método termina en 10 pasos.
El proceso termina, porque el menor AIC de las variables que quedan en el modelo deja de disminuir. La primera variable eliminada del modelo es rodilla, cuyo valor AIC =747.36 es el más pequeño de todos, luego se eliminan, pecho, altura, tobillo, biceps, en ese orden. El mejor modelo para predecir el porcentaje de grasa en el cuerpo será el que incluye a las variables: peso, edad, circunferencia de abdomen, nuñeca ,muslo,cadera,cuello y antebrazo.

El mejor modelo será:

Grasa= -22.65 -.089 peso+ .944 abdomen +0.515 antebrazo -1.53muñeca+0.065edad-0.466cuello-0.195cadera+.302muslo

El cual tiene un R^2 de 74.66, mientras que el modelo completo con 13 variable predictoras tiene un R^2 de 74.90%, se ha perdido un 0.24% de confiablidad en las predicciones pero se ha economizado 5 variables, lo cual es más conveniente.

Método de Selección hacia adelante

Aquí en el paso inicial se considera una regresión lineal simple que incluye a la variable predictora que da la correlación más alta con la variable de respuesta.

Se incluye una segunda variable en el modelo, que es aquella variable dentro de las no incluidas aún, mejora un criterio. Por ejemplo, da el "p-value" más bajo para la prueba *t* o el valor de la prueba de *t* más grande en valor absoluto. Y así se siguen incluyendo variables, notando que una vez que ésta es incluida ya no puede ser sacada del modelo.

El proceso termina cuando el criterio ya no puede ser mejorado. Por ejemplo, los "p-values" para la prueba *t* de todas las variables que aún no han sido incluidas son mayores que .05 ó la prueba de *t* es menor que 2 para dichas variables.

Ejemplo (cont). En el primer paso se halla la regresión simple con la variable predictora más altamente correlacionada con la variable de respuesta. En este caso, es *abdomen* que tiene correlación 0.803 con *grasa*.

La segunda variable que entra al modelo es peso porque es aquella con el valor de *t* más grande en valor absoluto entre las doce variables que aún no estaban incluidas.

La salida en R es como sigue:

```
> selforw(grasa[,2:14],grasa[,1],.15)
Seleccion Forward
```

p=numero de coeficientes en el modelo, p=1 es por el intercepto
nvar=p-1=numero de variables predictoras
add.var=la variable que ha sido anadida al modelo actual
pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

```
  p nvar   add.var  pvmax     s    r2 r2adj     Cp
2 2    1   abdomen 0.0000 4.877 0.662 0.660 72.869
3 3    2      peso 0.0000 4.456 0.719 0.717 20.691
4 4    3    muneca 0.0047 4.393 0.728 0.724 14.210
5 5    4 antebrazo 0.0098 4.343 0.735 0.731  9.314
6 6    5    cuello 0.1000 4.328 0.738 0.733  8.559
7 7    6      edad 0.1098 4.314 0.741 0.734  7.973
8 8    7     muslo 0.1098 4.291 0.744 0.737  6.338
```

La variable antebrazo es la ultiam en entra al modelo porque es aquella con el valor de *t* más grande en valor absoluto entre todas las variables que aún no estaban incluidas. Aquí termina el proceso porque al hacer las regresiones de grasa con las cuatro variables consideradas hasta ahora y cada una de las 9 variables no incluidas hasta ahora se obtienen “p-values” para la prueba *t* mayores de 0.05.

Regression Analysis

The regression equation is

grasa = - 34.9 + 0.996 abdomen - 0.136 peso - 1.51 muñeca + 0.473 antebrazo

Predictor	Coef	StDev	T	P
Constant	-34.854	7.245	-4.81	0.000
abdomen	0.99575	0.05607	17.76	0.000
peso	-0.13563	0.02475	-5.48	0.000
muñeca	-1.5056	0.4427	-3.40	0.001
antebraz	0.4729	0.1817	2.60	0.010

S = 4.343   R-Sq = 73.5%   R-Sq(adj) = 73.1%

Método Paso a Paso

Es una modificación del método "Forward", donde una variable que ha sido incluida en el modelo en un paso previo puede ser eliminada posteriormente.

En cada paso se cotejan si todas las variables que están en el modelo deben permanecer alli. La mayoría de las veces, pero no siempre, los tres métodos dan el mismo resultado para el mejor modelo de regresión.

En **MINITAB**, la opción ***Stepwise*** del submenú **Regression** selecciona el mejor modelo de regresión usando los métodos "**Stepwise".**

C) Usando el método "Stepwise".

sigue la secuencia

STAT ▸ **Regression** ▸ **Stepwise** ▸ **Methods** y luego se elige Stepwise. **Alpha-to-Enter** y **Alpha to-Remove.**

Para el conjunto de datos **grasa** el Método "stepwise" usa **Alpha-to-Enter** = 0.10 y **Alpha to-Remove** = 0.15.

El alpha to remove debe ser mayor o igual que el alpha to enter

Stepwise Regression: grasa versus edad, peso, ...

Alpha-to-Enter: 0.1 Alpha-to-Remove: 0.15

Response is grasa on 13 predictors, with N = 252

Step	1	2	3	4	5
Constant	-39.28	-45.95	-27.93	-34.85	-30.65
abdomen	0.631	0.990	0.975	0.996	1.008
T-Value	22.11	17.45	17.37	17.76	17.89
P-Value	0.000	0.000	0.000	0.000	0.000
peso		-0.148	-0.114	-0.136	-0.123
T-Value		-7.11	-4.84	-5.48	-4.75
P-Value		0.000	0.000	0.000	0.000
muneca			-1.24	-1.51	-1.25
T-Value			-2.85	-3.40	-2.66
P-Value			0.005	0.001	0.008
antebraz				0.47	0.53
T-Value				2.60	2.86
P-Value				0.010	0.005
cuello					-0.37
T-Value					-1.65
P-Value					0.100
S	4.88	4.46	4.39	4.34	4.33
R-Sq	66.17	71.88	72.77	73.50	73.79
R-Sq(adj)	66.03	71.65	72.44	73.07	73.26
C-p	72.9	20.7	14.2	9.3	8.6

- La primera variable seleccionada es abdomen, porque es la que tiene la prueba estadistica de t mas grande(o p-value mas pequeno). Es decir, abdomen es la mas importante para predecir el porcentaje de grasa. Las segunda variable mas importantes es peso, la tercera, muneca, la cuarta antebrazo y la quinta cuello. El metodo para en el paso 5 porque ninguna de las variables que aun no se han escogio som importantes para predcir grasa, es decir p-values debe ser mayor que el 10%(f-to-enter).
- Ademas en cada paso no se elimino ninguna variable que ya habia sido escogida previamente

Método de los mejores subconjuntos.

La opción **Best Subsets** del submenú **Regression** del menú **Stat** se usa para seleccionar los mejores modelos para un número dado de variables de acuerdo a 3 criterios:

El coeficiente de Determinación. El mejor modelo es aquel con R^2 más alto pero con el menor número de variables posibles.

$$R^2 = \frac{SSR}{SST}$$

El coeficiente de Determinación Ajustado. Es una variante del R^2 y que a diferencia de éste no aumenta necesariamente al incluir una variable adicional en el modelo.

$$R^2_{Ajust} = \frac{MSR}{MST} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

El **Coeficiente Cp de Mallows**. El mejor modelo es aquel para el cual se cumple aproximadamente , pero con $C_p = p+1$ el menor número de variables posibles. Notar que la igualdad anterior también se cumple cuando se usa el modelo completo.

$$C_p = \frac{SSE_p}{s^2} + 2(p+1) - n$$

Best Subsets Regression: grasa versus edad, peso, ...

Response is grasa

```
                                                              a
                                                              n
                                                  a     r t   t
                                            a c   b c   o o b e m
                                            l u p d a m d b i b u
                                        e p t e e o d u i i c r n
                                        d e u l c m e s l l e a e
                       Mallows          a s r l h e r l l l p z c
Vars  R-Sq  R-Sq(adj)      C-p       S  d o a o o n a o a o s o a
   1  66.2       66.0     72.9  4.8775            X
   2  71.9       71.7     20.7  4.4556    X       X
   3  72.8       72.4     14.2  4.3930    X       X             X
   4  73.5       73.1      9.3  4.3427    X       X           X X
   5  73.8       73.3      8.6  4.3276    X   X   X           X X
   6  74.1       73.5      7.7  4.3111  X X       X   X       X X
   7  74.4       73.7      6.3  4.2906  X X   X   X   X       X X
   8  74.7       73.8      6.4  4.2819  X X   X   X X X       X X
   9  74.8       73.8      7.2  4.2808  X X   X   X X X     X X X
  10  74.8       73.8      8.5  4.2832  X X   X   X X X   X X X X
  11  74.9       73.7     10.1  4.2879  X X X X   X X X   X X X X
  12  74.9       73.6     12.0  4.2963  X X X X X X X X   X X X X
  13  74.9       73.5     14.0  4.3053  X X X X X X X X X X X X X
```

Resultados para el problema anterior

De acuerdo al R^2 el mejor modelo podría ser aquel con las dos variables predictoras peso y abdomen que aún cundo su R^2 es de 71.9 está cerca del mayor posible que es de 74.9 y además es donde el R^2 ha tenido un mayor incremento. Un resultado similar cuando se usa el R^2 ajustado. De acuerdo al C_p de Mallows, el mejor modelo es aquel que tiene las siguientes 6 variables predictoras: edad, peso, muslo, abdomen, antebrazo y muneca con un valor de $C_p=7.7$ *muy próximo a* $p+1=7$.