Introduccción al aprendizaje estadístico

Ilustración regresión lineal: ventas vs gasto financiero

#### Juan David Ospina Arango

Técnicas en Aprendizaje Estadístico

Universidad Nacional de Colombia - Sede Medellín

#### 28 de mayo de 2019

Introducción

En este documento se ilustran algunos aspectos de un proceso de regresión lineal. La aplicación tiene que ver con predecir el gasto financiero de las empresas a partir de su nivel de ventas.

Lectura de los datos

La información que se utilizará son los reportes de ventas y de gasto financiero de las empresas para un año en particular. Esta información es pública y puede ser descargada desde el sitio web de la [Superintendencia de Sociedades](https://www.supersociedades.gov.co/delegatura_aec/estudios_financieros/Paginas/sirem.aspx):

ventas\_vs\_gastofinanciero<-read.csv("ventas\_vs\_gastofinanciero.csv",sep=";",  
 encoding="UTF-8")

¿Cuantos subsectores hay en la base de datos?

length(unique(ventas\_vs\_gastofinanciero$SUBSECTOR))

## [1] 84

¿Cuántos clientes tiene cada sector?

table(ventas\_vs\_gastofinanciero$SUBSECTOR)

##   
## ACUEDUCTO Y ALCANTARILLADO   
## 17   
## AGROQUÍMICOS   
## 13   
## ALIMENTOS CONCENTRADOS   
## 19   
## ARROZ   
## 5   
## ASEO   
## 11   
## AUTOMOTORES   
## 153   
## AVICULTURA   
## 32   
## AZÚCAR   
## 17   
## BANANO   
## 34   
## BANCOS   
## 1   
## BEBIDAS ALCOHÓLICAS   
## 8   
## BEBIDAS NO ALCOHÓLICAS   
## 7   
## CACAO, CHOCOLATE Y CONFITERÍA   
## 10   
## CAJAS DE COMPENSACIÓN   
## 6   
## CALZADO Y MARROQUINERÍA   
## 14   
## CARNES FRÍAS   
## 11   
## CEMENTO   
## 13   
## CEREALES   
## 3   
## COMBUSTIBLES Y LUBRICANTES   
## 28   
## COMERCIO DE CAFÉ   
## 40   
## COMERCIO DE ELECTRODOMÉSTICOS, MUEBLES Y COMPUTADORES   
## 42   
## COMERCIO DE FARMACÉUTICOS, COSMÉTICOS Y PRODUCTOS DE TOCADOR   
## 68   
## COMERCIO DE QUÍMICOS Y AGROQUÍMICOS   
## 39   
## COMERCIO DE REPUESTOS   
## 32   
## COMERCIO DE VARIEDADES Y VESTUARIO   
## 163   
## COMISIONISTAS DE BOLSA   
## 3   
## COMPANÍAS DE FINANCIAMIENTO COMERCIAL   
## 2   
## CONFECCIONES   
## 38   
## CONSERVAS, ALIÑOS Y SALSAS   
## 25   
## CURTIMBRES   
## 1   
## DEPARTAMENTOS   
## 2   
## DISTRIBUIDORES PRODUCTOS CONSUMO MASIVO   
## 153   
## EDIFICACIONES   
## 369   
## ELECTRICIDAD   
## 38   
## EPS Y SALUD PREPAGADA (SALUD)   
## 4   
## EQUIPO DE TRANSPORTE   
## 15   
## ESTABLECIMIENTOS EDUCATIVOS   
## 30   
## EXTRACCIÓN DE CARBÓN   
## 14   
## EXTRACCIÓN DE PETRÓLEO   
## 40   
## FERRETERÍA, MATERIALES DE CONSTRUCCIÓN Y MAQUINARÍA   
## 155   
## FIDUCIARIAS   
## 6   
## FLORES   
## 12   
## FONDOS DE PENSIONES   
## 2   
## GANADERÍA Y PORCICULTURA   
## 25   
## GAS   
## 15   
## GRASAS Y ACEITES   
## 18   
## HOTELES, TURISMO Y AGENCIAS DE TURISMO   
## 39   
## IMPRENTAS Y EDICIÓN   
## 24   
## INDUSTRIA BÁSICA DE HIERRO Y ACERO   
## 14   
## IPS   
## 80   
## MADERA Y MUEBLES   
## 22   
## MANUFACTURA DE PRODUCTOS DE TOCADOR Y ARTICULOS DE ASEO   
## 18   
## MANUFACTURA DE PRODUCTOS FARMACÉUTICOS   
## 39   
## MAQUINARIA   
## 47   
## MEDIOS   
## 42   
## MOLINERÍA Y DERIVADOS   
## 37   
## OBRAS DE INFRAESTRUCTURA   
## 204   
## OTRAS ACTIVIDADES MINERAS   
## 8   
## OTROS CULTIVOS   
## 23   
## OTROS MANUFACTURA   
## 15   
## OTROS SERVICIOS FINANCIEROS   
## 62   
## PALMA AFRICANA   
## 27   
## PAPEL Y CARTÓN   
## 30   
## PERSONAS   
## 5   
## PESCA Y ATÚN   
## 3   
## PINTURAS Y BARNICES   
## 9   
## PLÁSTICO   
## 63   
## PRODUCTOS CERÁMICOS   
## 12   
## PRODUCTOS LÁCTEOS   
## 18   
## PRODUCTOS METÁLICOS   
## 51   
## QUÍMICOS BÁSICOS Y OTROS QUÍMICOS   
## 42   
## REPUESTOS Y CAUCHO   
## 24   
## RESINAS PLÁSTICAS   
## 6   
## SEGUROS   
## 3   
## SERVICIOS A EMPRESAS   
## 303   
## SERVICIOS A PERSONAS   
## 216   
## SUPERMERCADOS   
## 48   
## TABACO   
## 1   
## TELCOS   
## 12   
## TEXTILES   
## 36   
## TRANSPORTE AEREO   
## 13   
## TRANSPORTE MARÍTIMO   
## 25   
## TRANSPORTE TERRESTRE   
## 130   
## VIDRIO   
## 9

Esto podría verse mejor como un dataframe:

sectores<-as.data.frame(table(ventas\_vs\_gastofinanciero$SUBSECTOR))  
head(sectores)

Ahora cambiemos los nombres de las variables para mejorar la comprensión del conjunto de datos:

names(sectores)<-c("subsector","cant\_clientes")  
head(sectores)

Ahora veamos los subsectores con más clientes:

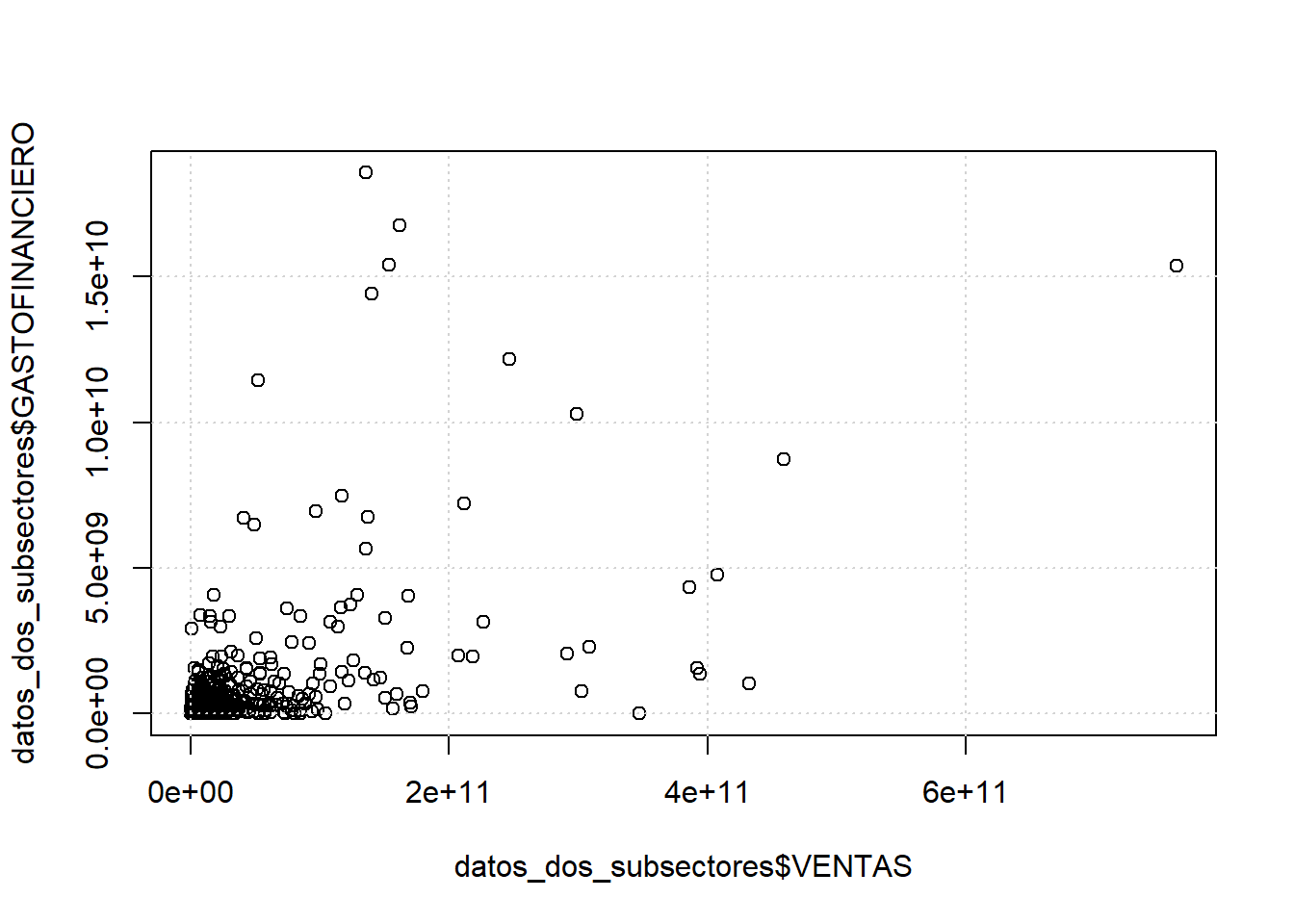
sectores[order(sectores$cant\_clientes,decreasing=TRUE),]

Vamos a escoger dos subsectores y a visualizar la relación entre sus ventas y el gasto financiero:

ids<-ventas\_vs\_gastofinanciero$SUBSECTOR %in% c("EDIFICACIONES","COMERCIO DE VARIEDADES Y VESTUARIO")  
datos\_dos\_subsectores<-ventas\_vs\_gastofinanciero[ids,]

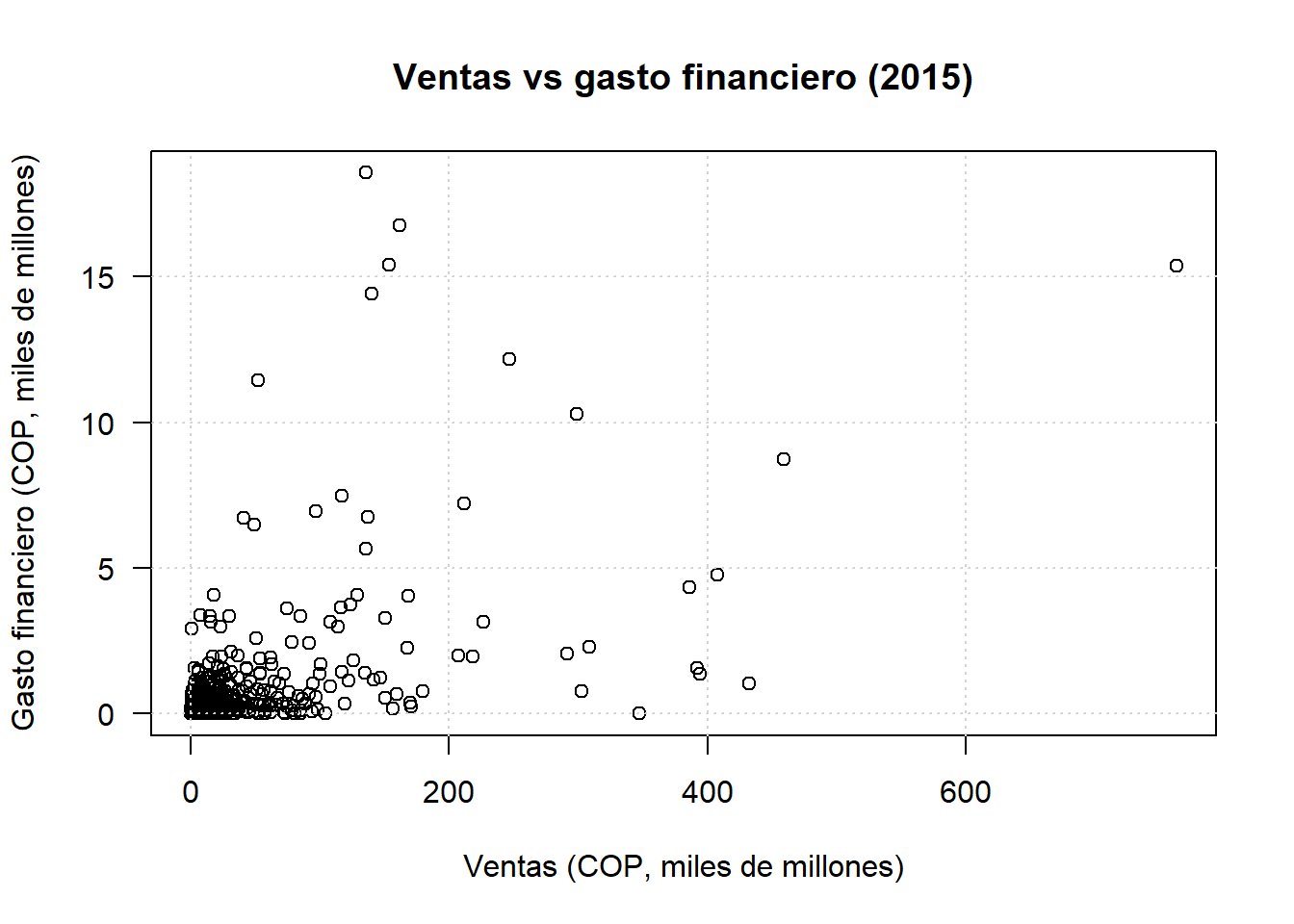
Ahora grafiquemos los datos con la función *plot()*:

plot(datos\_dos\_subsectores$VENTAS,datos\_dos\_subsectores$GASTOFINANCIERO)  
grid()



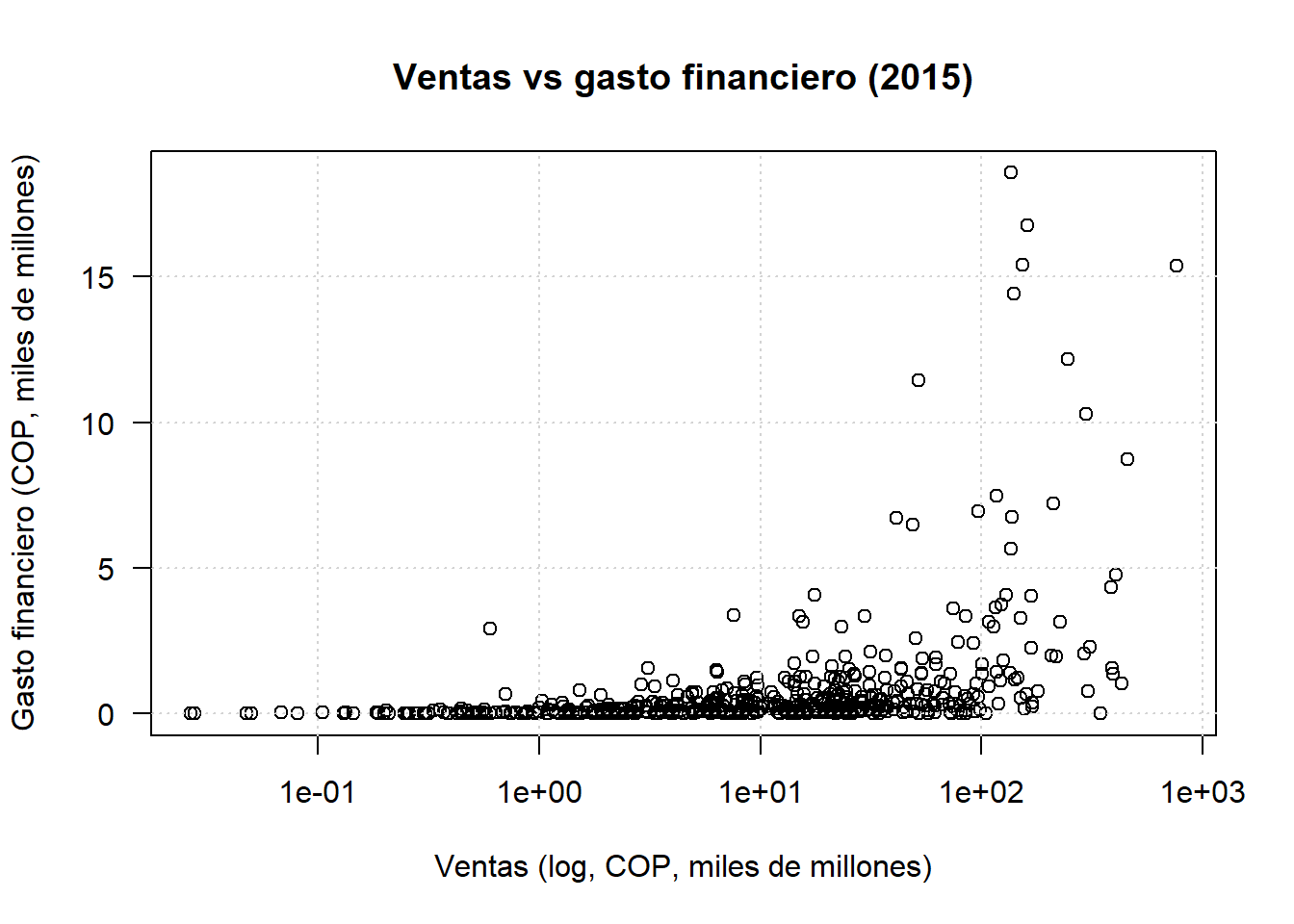
Esta gráfica tiene problemas de presentación. El siguiente código muesta algunas ideas para mejorarla:

plot(datos\_dos\_subsectores$VENTAS/10e8,datos\_dos\_subsectores$GASTOFINANCIERO/10e8,  
 xlab="Ventas (COP, miles de millones)",ylab = "Gasto financiero (COP, miles de millones)",main="Ventas vs gasto financiero (2015)",  
 las=1)  
grid()



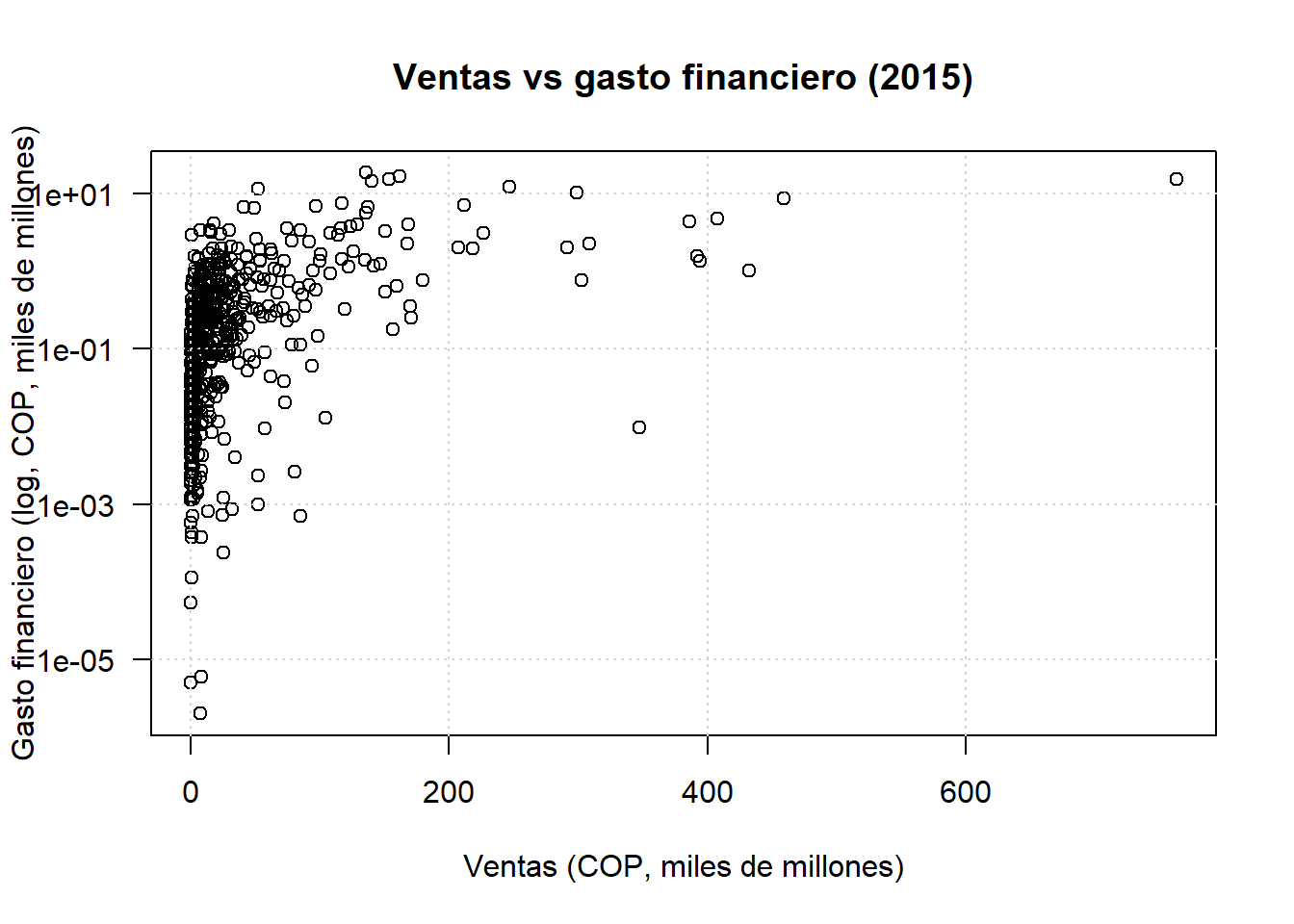
Veamos que pasa cuando tenemos las ventas en escala logarítmica:

plot(datos\_dos\_subsectores$VENTAS/10e8,datos\_dos\_subsectores$GASTOFINANCIERO/10e8,  
 xlab="Ventas (log, COP, miles de millones)",ylab = "Gasto financiero (COP, miles de millones)",main="Ventas vs gasto financiero (2015)",  
 las=1,log="x")  
grid()



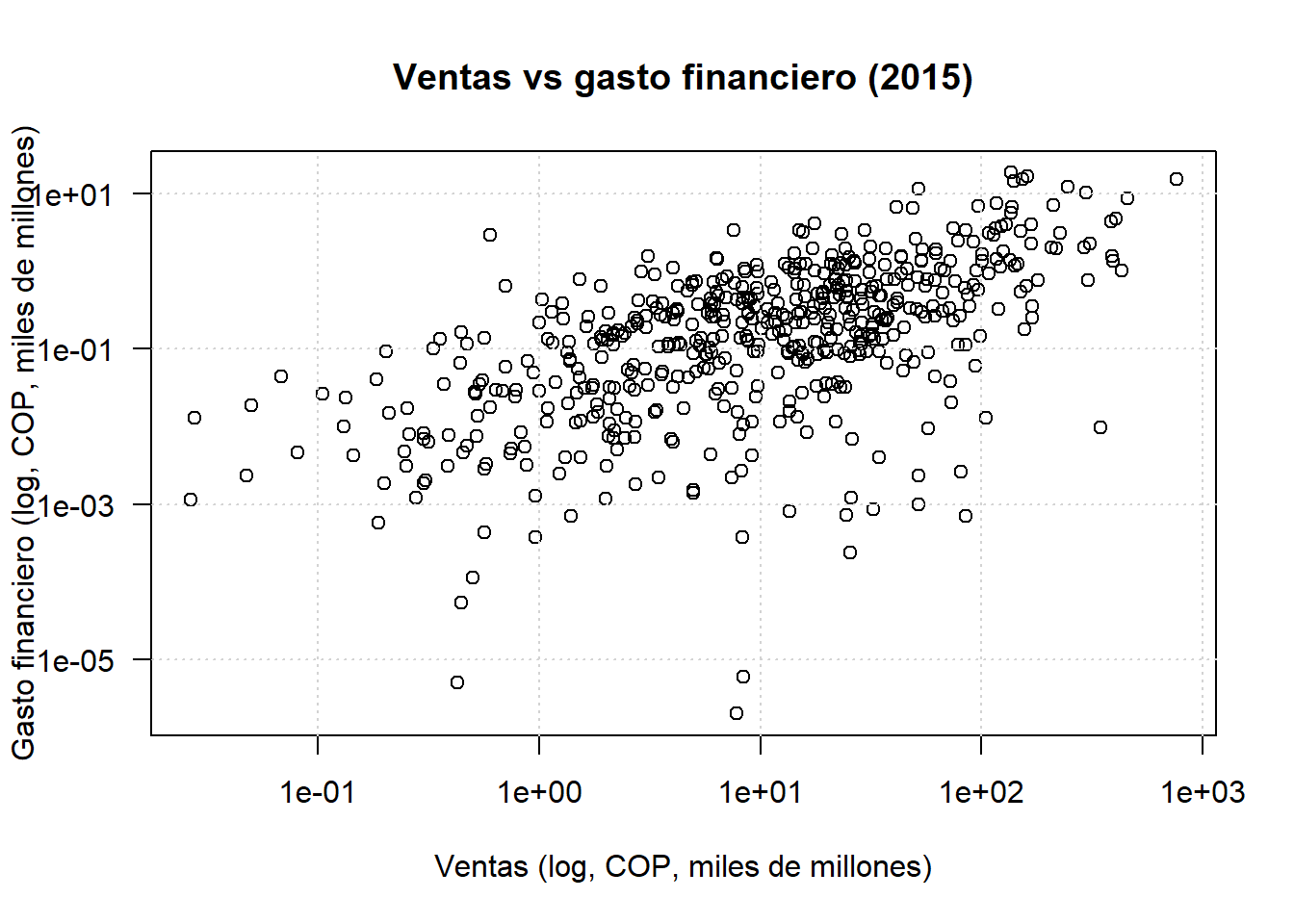
Veamos que pasa cuando tenemos el gasto financiero en escala logarítmica:

plot(datos\_dos\_subsectores$VENTAS/10e8,datos\_dos\_subsectores$GASTOFINANCIERO/10e8,  
 xlab="Ventas (COP, miles de millones)",ylab = "Gasto financiero (log, COP, miles de millones)",main="Ventas vs gasto financiero (2015)",  
 las=1,log="y")  
grid()



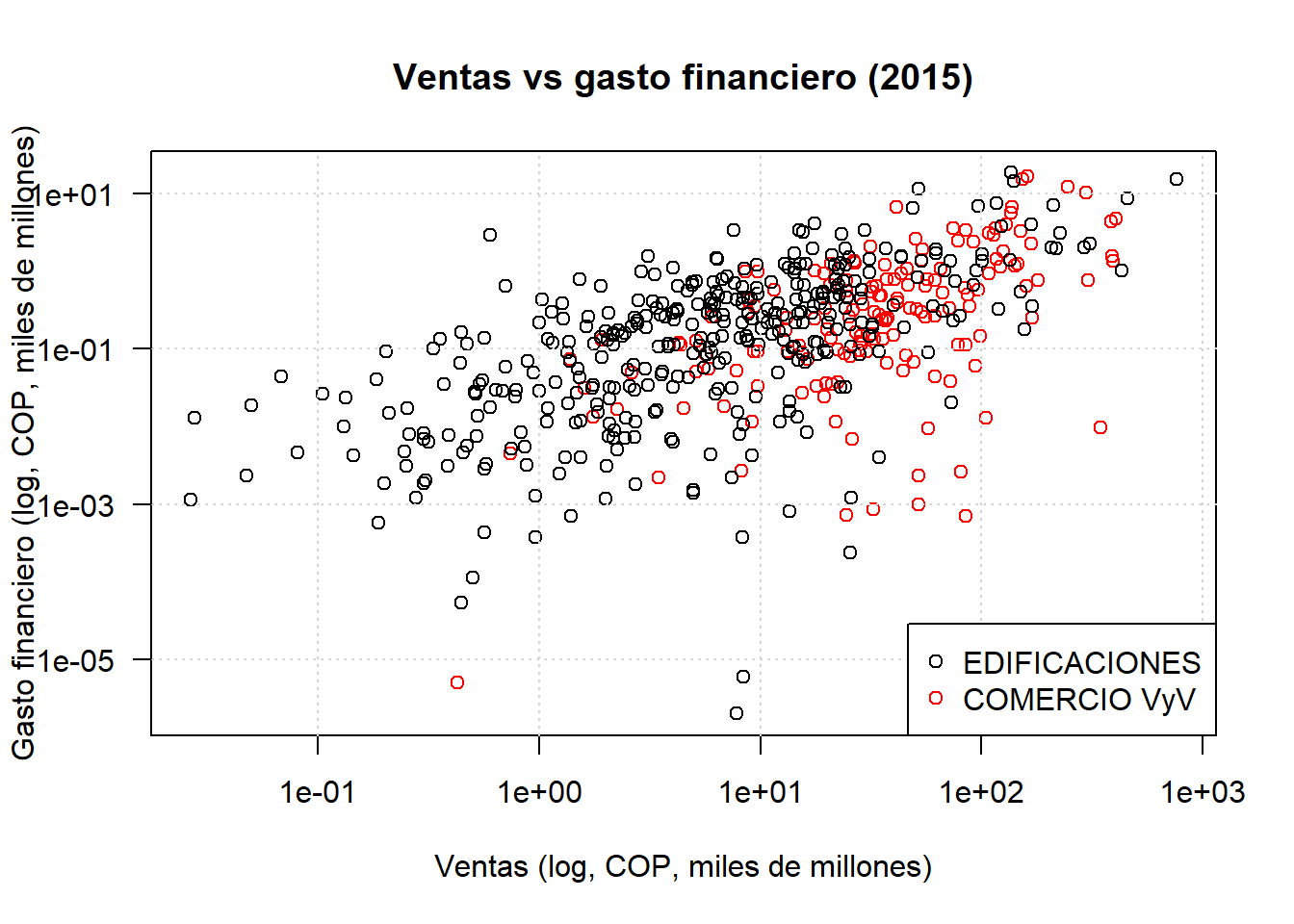
Veamos que pasa cuando tenemos ambos,las ventas y el gasto financiero, en escala logarítmica:

plot(datos\_dos\_subsectores$VENTAS/10e8,datos\_dos\_subsectores$GASTOFINANCIERO/10e8,  
 xlab="Ventas (log, COP, miles de millones)",ylab = "Gasto financiero (log, COP, miles de millones)",main="Ventas vs gasto financiero (2015)",  
 las=1,log="xy")  
grid()



Aunque ahora pareciera haber una relación lineal, ésta podría no ser la misma para los dos subsectores en cuestión.

colores<-ifelse(datos\_dos\_subsectores$SUBSECTOR=="EDIFICACIONES",1,2)  
plot(datos\_dos\_subsectores$VENTAS/10e8,datos\_dos\_subsectores$GASTOFINANCIERO/10e8,  
 xlab="Ventas (log, COP, miles de millones)",ylab = "Gasto financiero (log, COP, miles de millones)",main="Ventas vs gasto financiero (2015)",  
 las=1,log="xy",col=colores)  
grid()  
legend("bottomright",legend=c("EDIFICACIONES","COMERCIO VyV"),col=c(1,2),pch=1)



Probemos un modelo lineal para el sector de edificaciones. Para ello transformemos los datos primero:

datos\_dos\_subsectores\_trs<-datos\_dos\_subsectores  
datos\_dos\_subsectores\_trs$VENTAS<-log(datos\_dos\_subsectores\_trs$VENTAS/10e8)  
datos\_dos\_subsectores\_trs$GASTOFINANCIERO<-log(datos\_dos\_subsectores\_trs$GASTOFINANCIERO/10e8)

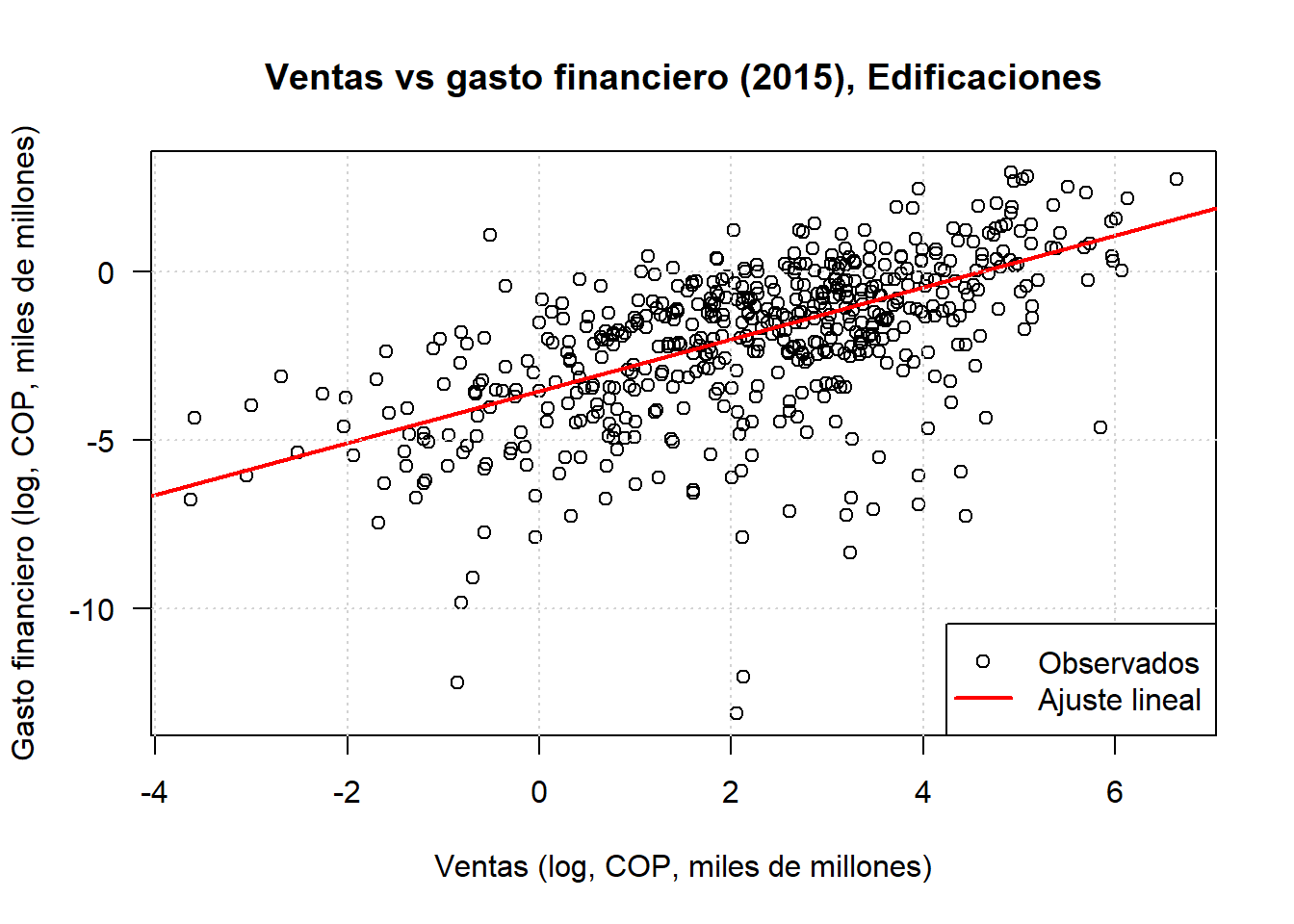
Ahora ajustemos un modelo para edificaciones:

modelo\_ed\_1<-lm(GASTOFINANCIERO~VENTAS,data=datos\_dos\_subsectores\_trs,  
 subset=(SUBSECTOR=="EDIFICACIONES"))  
summary(modelo\_ed\_1)

##   
## Call:  
## lm(formula = GASTOFINANCIERO ~ VENTAS, data = datos\_dos\_subsectores\_trs,   
## subset = (SUBSECTOR == "EDIFICACIONES"))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.1541 -0.8871 0.3045 1.2541 5.0192   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.55769 0.13670 -26.02 <2e-16 \*\*\*  
## VENTAS 0.77140 0.05541 13.92 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.911 on 367 degrees of freedom  
## Multiple R-squared: 0.3456, Adjusted R-squared: 0.3438   
## F-statistic: 193.8 on 1 and 367 DF, p-value: < 2.2e-16

Veamos el ajuste del modelo en una gráfica:

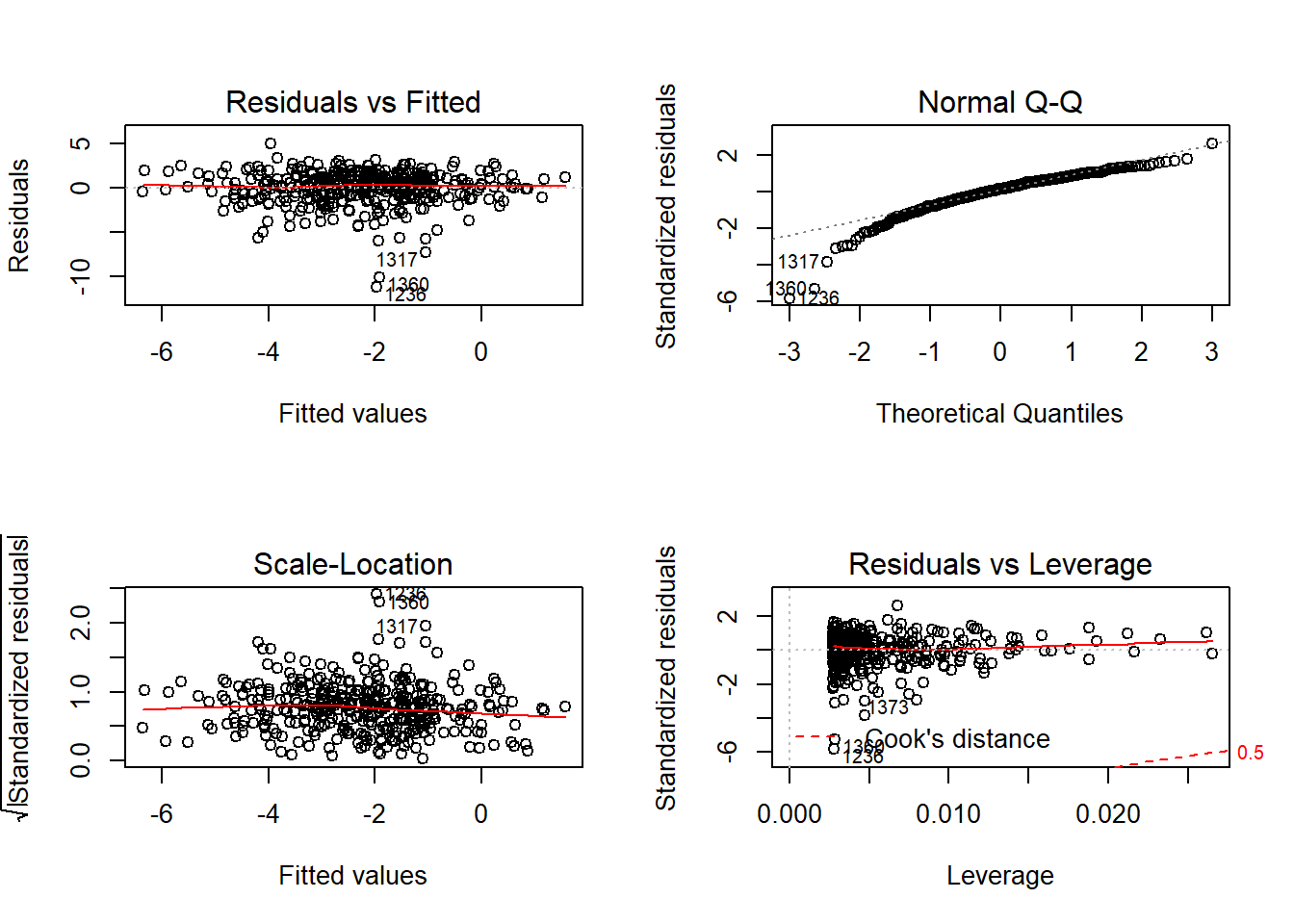
plot(datos\_dos\_subsectores\_trs$VENTAS,datos\_dos\_subsectores\_trs$GASTOFINANCIERO,  
 xlab="Ventas (log, COP, miles de millones)",ylab = "Gasto financiero (log, COP, miles de millones)",main="Ventas vs gasto financiero (2015), Edificaciones",  
 las=1)  
abline(modelo\_ed\_1,col="red",lwd=2)  
grid()  
legend("bottomright",legend = c("Observados","Ajuste lineal"),pch=c(1,NA),lty=c(NA,1),  
 col=c(1,2),lwd=c(1,2))



Diagnóstico: tarea

Se recomienda investigar y entender el diagnóstico del modelo de regresión lineal para comprender y utilizar las siguientes gráficas.

par(mfrow=c(2,2))  
plot(modelo\_ed\_1)



Actividades sugeridas

Se sugieren las siguientes actividades para afianzar las capacidades:

1. Para el modelo para edificaciones retirar los clientes que más se alejan del modelo, reestimar el modelo y observar el cambio en el error residual y en el \(R^2\). ¿Para cuántos clientes funciona el modelo? ¿Qué porcentaje de los clientes representan aquellos para los que no funciona el modelo?
2. Estimar un modelo para variedades y vestuario. ¿Qué tan similar es este modelo al de edificaciones?
3. Estime un modelo conjunto para edificaciones y para variedades y vestuario. ¿Qué tan similar es éste a los modelos individuales? ¿Es mejor estimar un modelo conjunto o dos individuales?
4. Predicciones: ¿cuál es el gasto financiero esperado para una empresa del sector de edificaciones con unas ventas de COP $400 mil millones? ¿Cuál es el intervalo de confianza para esta estimación? ¿Cuál sería el valor esperado y el intervalo de confianza para una empresa del sector de variedades y vestuarios con el mismo nivel de ventas?

FIN