Estadística Aplicada III

Análisis exploratorio de datos (cont)
Visualización

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana2: 22/24 de agosto de 2018

Análisis Exploratorio de datos

Ejemplo 1: Pena de muerte y la paradoja de Simpson I

- Las estructuras de dependencia en conjuntos de datos multivariados puede ser compleja.
- Un ejemplo interesante surge cuando se ignoran variables importantes que están relacionadas con la variable de interés.
- Consideren un conjunto de variables categóricas X, Y, Z, donde se consideran a Y como una variable de respuesta con m niveles, X es una variable explicativa con n niveles y Z es una variable adicional con l niveles.
- Una *tabla de contingencia de* $m \times n \times l$ muestra los conteos observados para las combinaciones de las tres variables.

Ejemplo 1: Pena de muerte y la paradoja de Simpson II

Ejemplo: Datos sobre pena de muerte (Radelet, 1991)

- Este estudio evalúa los efectos de las características raciales en la decisión de un juez de sentenciar a pena de muerte al ofensor.
- 674 fueron considerados en 20 condados de Florida entre 1976 y 1987. La tabla de contingencia de $2 \times 2 \times 2$ se muestra a continuación:

Raza de la víctima (Z)	Raza del ofensor (X)	Pena de muerte (Y) Si	No	% Pena de muerte
Blanca	Blanca	53	414	11.3
	Negra	11	37	22.9
Negra	Blanca	0	16	0
	Negra	4	139	2.8
Total	Blanca	53	430	11.0
	Negra	15	176	7.9

- Notar que cuando las víctimas son blancas, la pena de muerte se impuso $22.9\,\% - 11.3\,\% = 11.6\,\%$ más frecuente para negros que para blancos, y cuando la víctima es negra $2.8\,\%$ más para negros. Estas cantidades corresponden a las condicionales $Y|X, Z = 0$ y $Y|X, Z = 1$.
- Entonces, controlando por raza de la víctima, el porcentaje de sentencias a muerte es mayor para ofensores negros que para blancos.

Ejemplo 1: Pena de muerte y la paradoja de Simpson III

- Sin embargo, si se ignora la raza de la víctima (parte baja de la tabla, correspondiente a la marginal $Y|X$) entonces los ofensores blancos reciben $11\,\% - 7.9\,\% = 3.1\,\%$ más veces la pena de muerte. La asociación se invierte.
- Esta inversión entre tipos de asociación (asociación condicional vs. asociación marginal) se conoce como la *Paradoja de Simpson* o *efecto Yule-Simpson*. También se puede presentar en datos continuos.

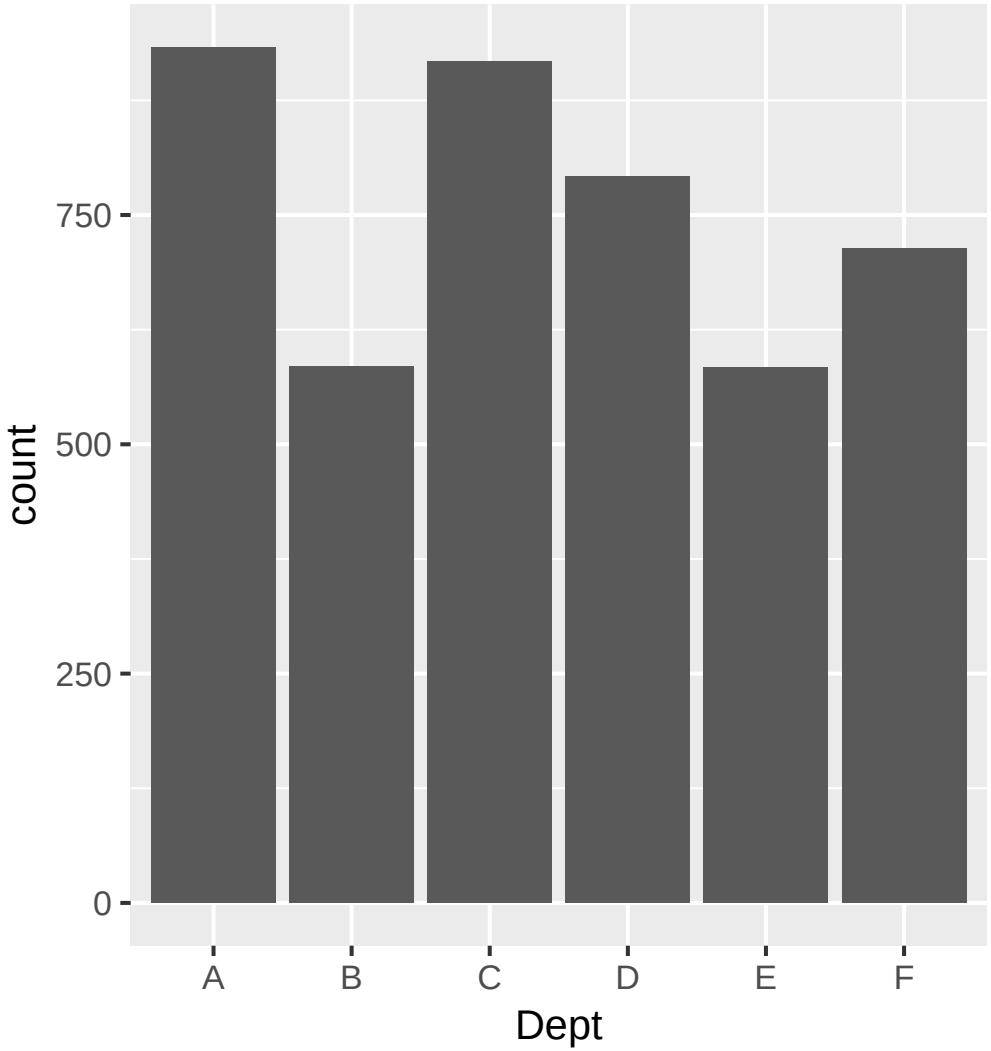
Ejemplo 2: Admisiones a UBerkeley I

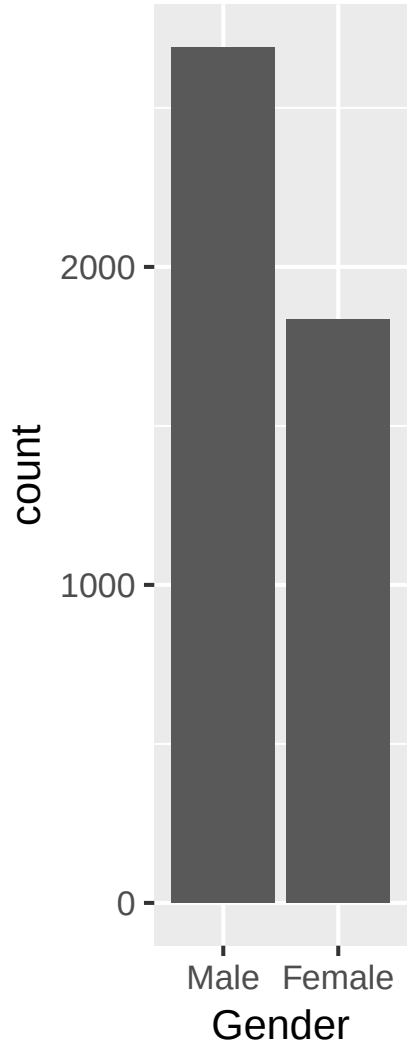
Ejemplo 2: Datos de admisiones de la Universidad de Berkeley (1973)

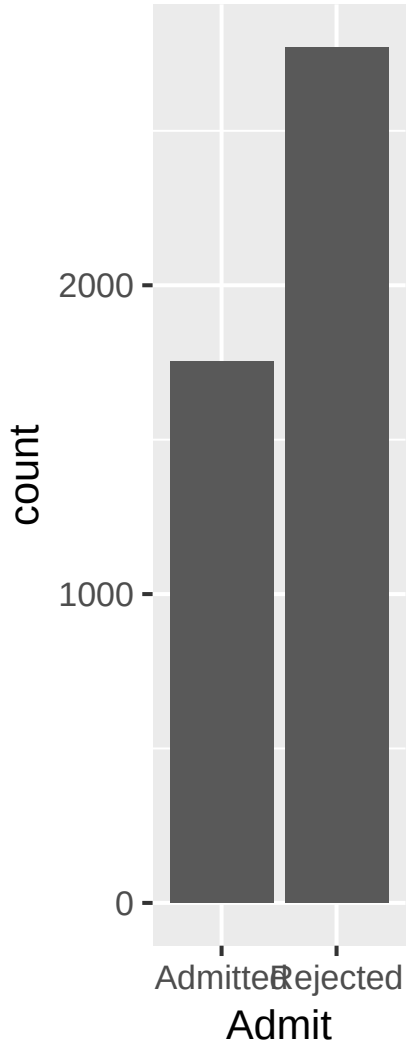
- Los datos se refieren a las solicitudes de acceso a posgrado de la Universidad de Berkeley en 1973 clasificadas por admisión y género.
- Una de las razones del estudio era determinar si había sesgo en las admisiones por el género del solicitante
- los datos son parte de los datos básicos de `R`: `UCBAdmissions`.
- las siguientes gráficas muestran las variables sin relacionarlas, de las cuales no se pueden hacer muchas inferencias.

```
suppressMessages(library(ggplot2)) #libreria de Hadley Wickham basado en grammar of graphics
suppressMessages(library(gridExtra)) #paquete para hacer arreglos de gráficas generadas con ggplot2
admisiones <- as.data.frame(UCBAdmissions)
a <- ggplot(admisiones, aes(Dept)) + geom_bar(aes(weight=Freq))
b <- ggplot(admisiones, aes(Gender)) + geom_bar(aes(weight=Freq))
c <- ggplot(admisiones, aes(Admit)) + geom_bar(aes(weight=Freq))
grid.arrange(a, b, c, nrow=1,widths=c(7,3,3))
```

Ejemplo 2: Admisiones a UBerkeley II







Ejemplo 2: Admisiones a UBerkeley III

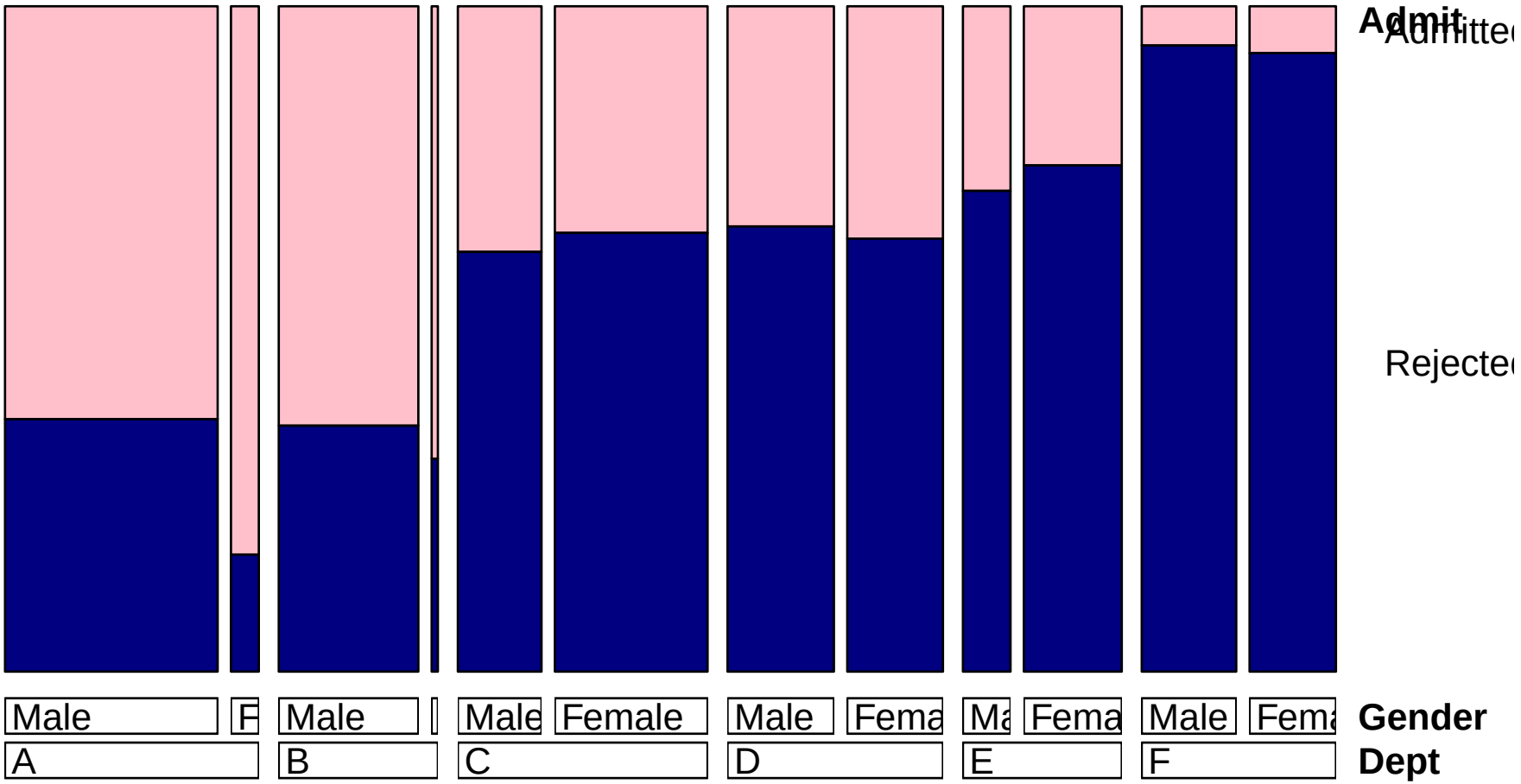
- Observando las tablas de contingencia, ignorando los departamentos, podemos ver que la tasa de aceptación para mujeres es sólo de cerca del 30 % y para los hombres es de cerca de 45 %:

```
a <- xtabs(Freq ~ Gender + Admit, data=admisiones)
a[1,1]/sum(a[1,]) #tasa de aceptación de hombres

[1] 0.4451877

a[2,1]/sum(a[2,]) #tasa de aceptación de mujeres

[1] 0.3035422
```

- Sin embargo, considerando la variable `Dept` para el departamento de admisión, se puede ver que en cuatro de los seis departamentos las mujeres tienen una mayor tasa de admisión.

```
library(vcd) #librería para gráficas de datos categóricos

Loading required package: grid

doubledecker(xtabs(Freq ~ Dept + Gender + Admit, data = admisiones),
        gp = gpar(fill=c("pink","navy")))
```

Ejemplo 2: Admisiones a UBerkeley IV



- La gráfica anterior es una *gráfica de mosaico*. En este ejemplo, el ancho de las barras son proporcionales a los números en los respectivos grupos.

Ejemplo 2: Admisiones a UBerkeley V

- En la gráfica, se puede ver que los departamentos A, B, D y F aceptaron más mujeres, y menos mujeres aplicaron a los departamentos A y B, y menos hombres en el C y E.
- Pero la gráfica no es suficiente: cualquier resultado que se encuentre gráficamente debe verificarse con métodos estadísticos. Más adelante revisaremos los modelos estadísticos para probar estos resultados gráficos.

Ejemplo 3: Resultados del examen *Suite of Assessments (SAT)* I

Ejemplo 2: Datos de admisiones de la Universidad de Berkeley (1973)

- Los datos corresponden a los resultados del examen SAT (score total promedio) de 1997 de escuelas en todos los estados de la Unión Americana y algunas variables como el salario promedio anual de los maestros.
- Se encontró que había una relacion negativa entre el salario promedio anual de los maestros y el score total promedio de los alumnos que presentaron el SAT. ¿Esto implicaba que habría que pagar menos a los maestros?

```
sat <- read.csv("~/Dropbox/Academia/ITAM/EA3S18-II/data/sat.csv")
head(sat)
```

```
       state expenditure pupil_teacher_ratio teacher_salary perc_take_sat verbal_score math_score total_score
1    Alabama       4.405                17.2         31.144             8          491        538        1029
2     Alaska       8.963                17.6         47.951            47          445        489         934
3    Arizona       4.778                19.3         32.175            27          448        496         944
4   Arkansas       4.459                17.1         28.934             6          482        523        1005
5 California       4.992                24.0         41.078            45          417        485         902
6   Colorado       5.443                18.4         34.571            29          462        518         980
```

Ejemplo 3: Resultados del examen *Suite of Assessments (SAT)* II

```r
plot(sat$teacher_salary,  sat$total_score)
abline(lm(total_score ~ teacher_salary,data=sat),lwd=2)
porcentaje <- cut(sat$perc_take_sat,breaks=c(0,20,50,81)) #genera un factor por rango de porcentaje
porcentaje[1:5]
```

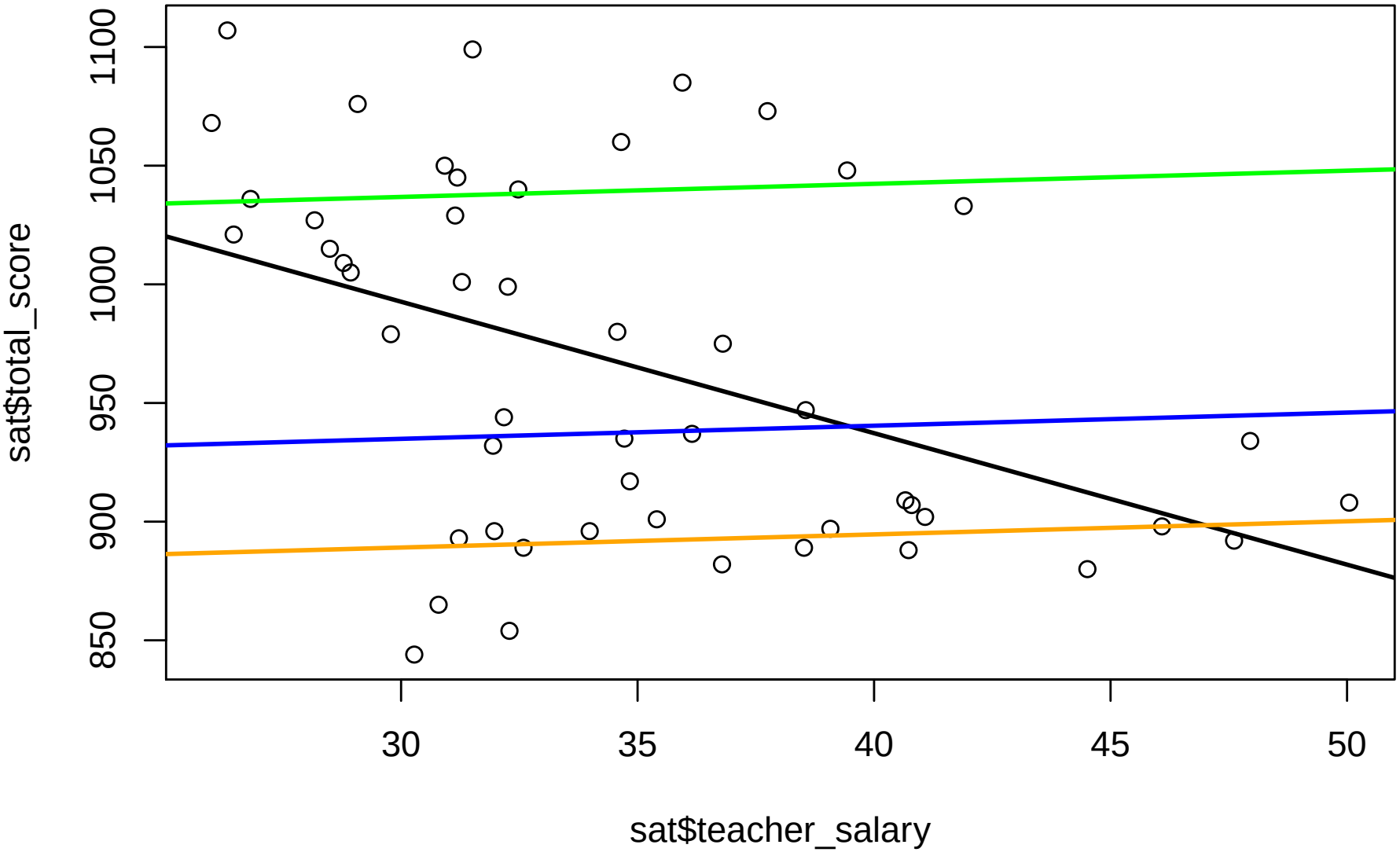
```
[1] (0,20]  (20,50] (20,50] (0,20]  (20,50]
Levels: (0,20] (20,50] (50,81]
```

```r
reg <- lm(total_score ~ teacher_salary + porcentaje, data=sat)
summary(reg)
```

```
Call:
lm(formula = total_score ~ teacher_salary + porcentaje, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-105.89  -19.70    0.48   13.68   74.78

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1020.2247    33.0599  30.860  < 2e-16 ***
teacher_salary       0.5532     1.0218   0.541    0.591
porcentaje(20,50] -101.9267    14.3323  -7.112 6.20e-09 ***
porcentaje(50,81] -147.7388    13.4755 -10.964 2.01e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.91 on 46 degrees of freedom
Multiple R-squared:  0.7838,Adjusted R-squared:  0.7697
F-statistic: 55.59 on 3 and 46 DF,  p-value: 2.461e-15
```

```r
abline(reg$coef[1:2], col="green",lwd=2)
abline(reg$coef[1] + reg$coef[3], reg$coef[2], col = "blue",lwd=2)
abline(reg$coef[1] + reg$coef[4], reg$coef[2], col = "orange",lwd=2)
```

Ejemplo 3: Resultados del examen *Suite of Assessments (SAT)* III

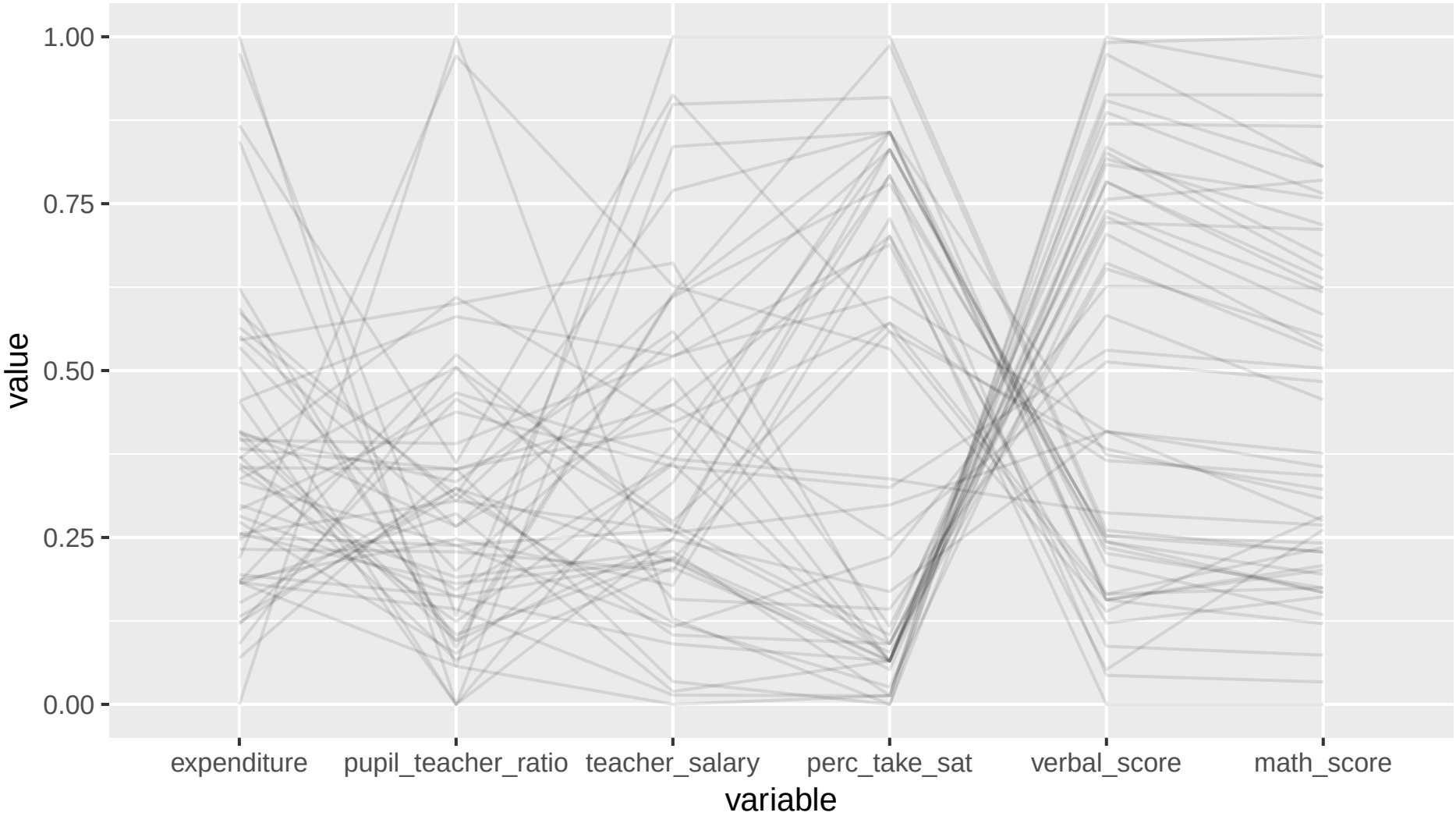


Gráficas de coordenadas paralelas I

- Las coordenadas asociadas a las dimensiones son paralelas unas a otras. Cada variable tiene su eje vertical y los ejes se escalan para que los rangos coincidan
- Los valores para cada caso se unen. Con esto se pueden ver los perfiles de cada caso y se puede ver si los datos forman clusters o agrupaciones

```
library(GGally)
ggparcoord(data=sat,columns=c(2:7),scale="uniminmax",alphaLines=0.1)
```

Gráficas de coordenadas paralelas II



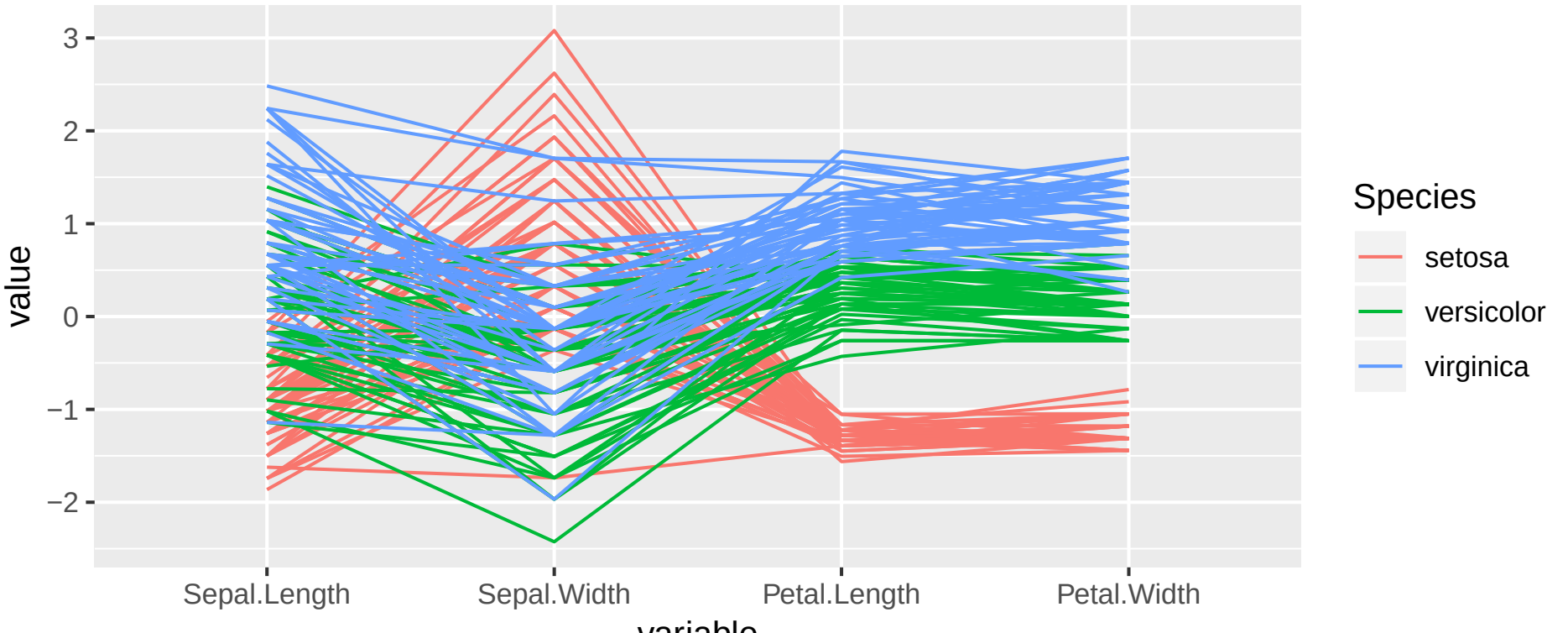
Gráficas de coordenadas paralelas

¿Qué se puede ver en una gráfica de coordenadas paralelas? Se pueden considerar diferentes ordenamientos de los ejes para observar todas las posibles adyacencias.

- ua visión general de la distribución unnivariada de cada variable
- agrupaciones
- correlaciones obvias
- asociaciones bivariadas en datos adyacentes.

Otro ejemplo:

```
ggparcoord(iris,columns=1:4,groupColumn="Species")
```



Gráficas dinámicas

- Existen una variedad de paquetes que permiten considerar gráficas dinámicas.
- Mondrian
- Ggobi
- Xlisp-Stat (vista)
- `ggvis`
- `iplots`
- ?
- Ver algunos ejemplos de uso prácticos. Durante el curso veremos las aplicaciones específicas.