

Regresión Avanzada

Variables cualitativas e interacciones

Variables Cualitativas

En muchas situaciones las variables no pueden estar bien definidas en una escala numérica.

Ejemplos:

- ▶ sexo
- ▶ nacionalidad
- ▶ marca
- ▶ estatus laboral

En muchos casos indican la presencia/ausencia de un atributo.

- ▶ empleado/desempleado
- ▶ fumador/no fumador
- ▶ graduado universitario / no graduado universitario

Variables Cualitativas En estos casos se pueden codificar con 0's y 1's.

- ▶ 1 indica la presencia del atributo.
- ▶ 0 indica la ausencia del atributo.

Estas variables se conocen como **variables indicadoras** o **dummy**.
Ejemplo:

$$D = \begin{cases} 1 & \text{la persona esta empleada} \\ 0 & \text{la persona esta desempleada} \end{cases}$$

Indistintamente se podría haber codificado

$$D = \begin{cases} 1 & \text{la persona esta desempleada} \\ 0 & \text{la persona esta empleada} \end{cases}$$

Conviene hacer la codificación que resulte más sencilla de interpretar a la hora de interpretar el problema.

Cuando las variables son

- ▶ **cuantitativas** se llama **modelo de regresión**.
- ▶ **cualitativas** se llama **modelo de análisis de varianza**.
- ▶ **cuantitativas y cualitativas** se llama **modelo de análisis de covarianza**.

Ejemplo

Consideramos el siguiente modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \epsilon,$$

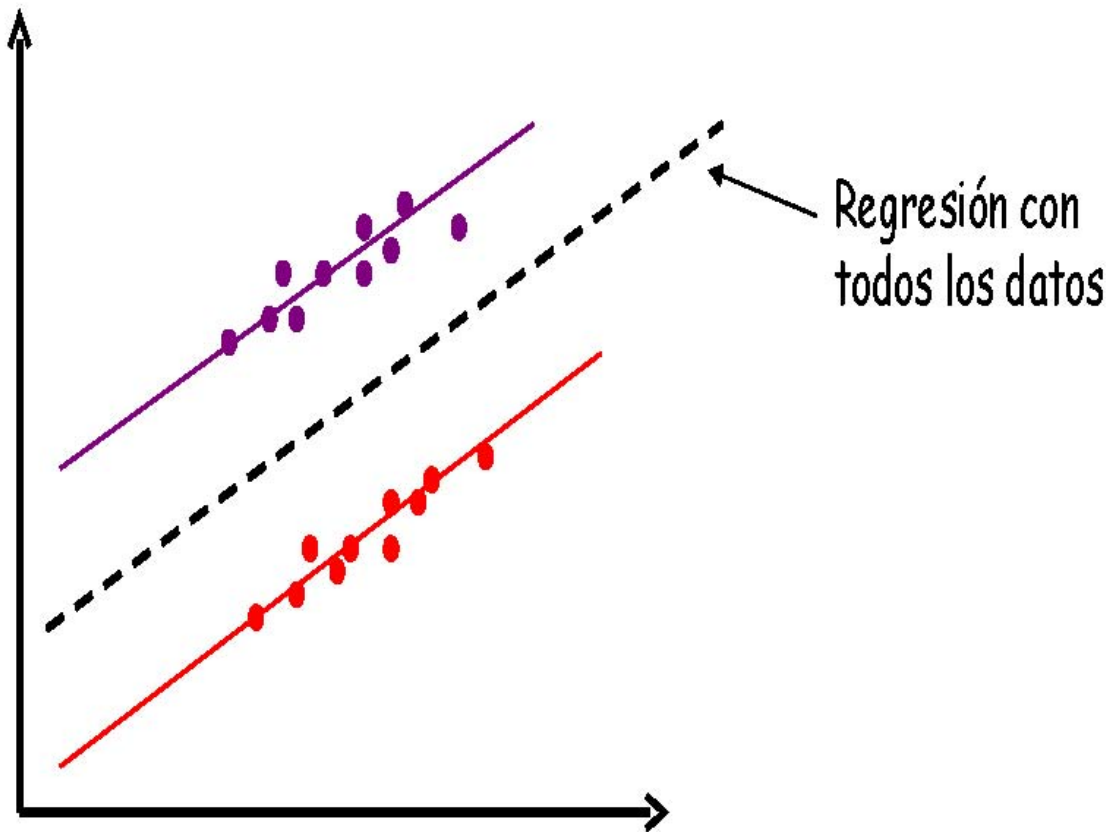
donde $E(\epsilon) = 0$ y $var(\epsilon) = \sigma^2$, además sabemos que la variable x_1 es cuantitativa, mientras que la variable D_2 es dummy,

$$D = \begin{cases} 1 & \text{la observación pertenece al grupo A} \\ 0 & \text{la observación pertenece al grupo B} \end{cases}$$

Luego tenemos que

- ▶ si $D_2 = 0$ entonces $y = \beta_0 + \beta_1 x_1 + \epsilon$,
- ▶ si $D_2 = 1$ entonces $y = \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon$.

Interpretación gráfica



Observación:
si no se pone la
variable dummy no
ajusta adecuadamente
a ninguna de las
dos poblaciones.

Es decir,

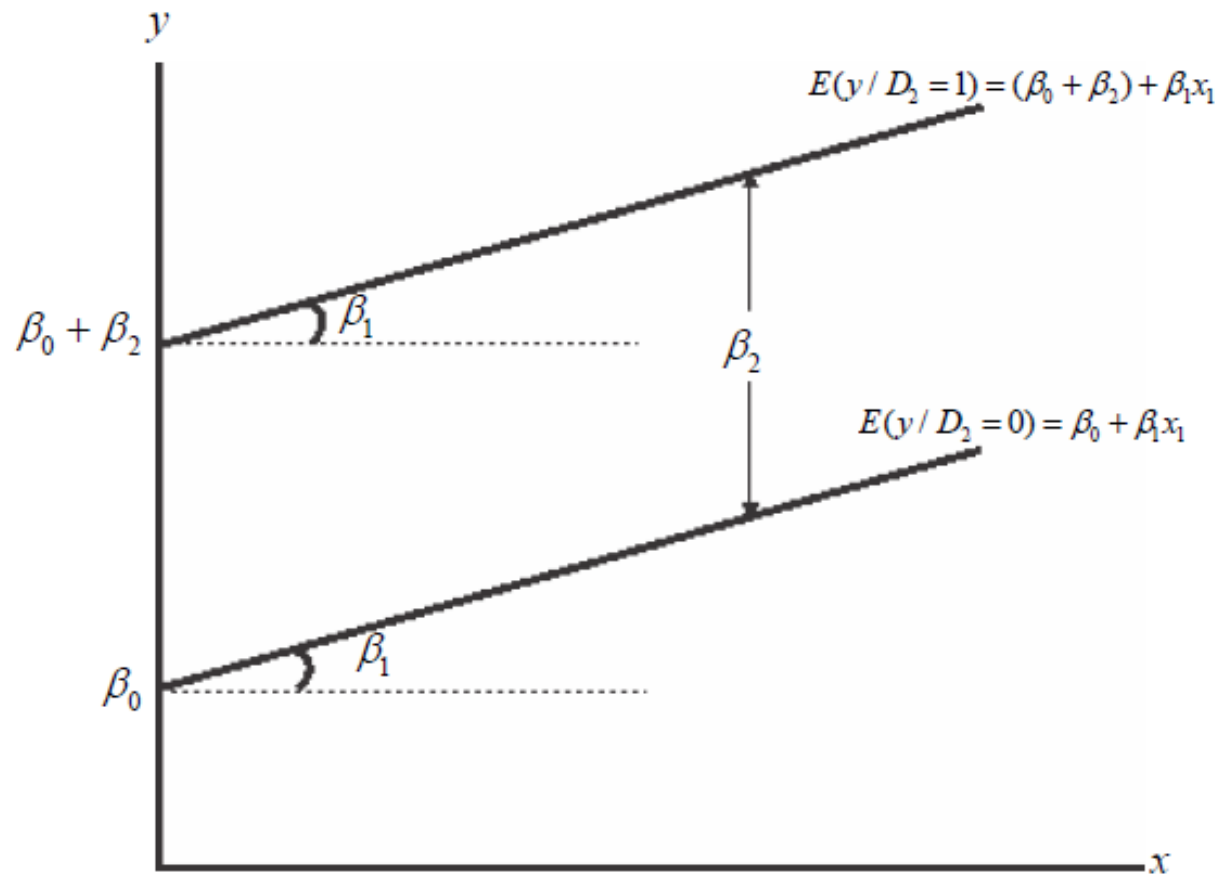
- ▶ si $E(y|D_2 = 0) = \beta_0 + \beta_1 x_1$, es una recta con pendiente β_1 e intercept β_0 .
- ▶ si $E(y|D_2 = 1) = \beta_0 + \beta_1 x_1 + \beta_2$, es una recta con pendiente β_1 e intercept $\beta_0 + \beta_2$.

Son dos rectas paralelas con distinto intercept. Luego, las cantidades $E(y|D_2 = 0)$ y $E(y|D_2 = 1)$ son las respuestas medias cuando las observaciones pertenecen al grupo A o al grupo B, y

$$\beta_2 = E(y|D_2 = 1) - E(y|D_2 = 0),$$

es la diferencia entre los valores medios de y cuando $D = 1$ y cuando $D = 0$.

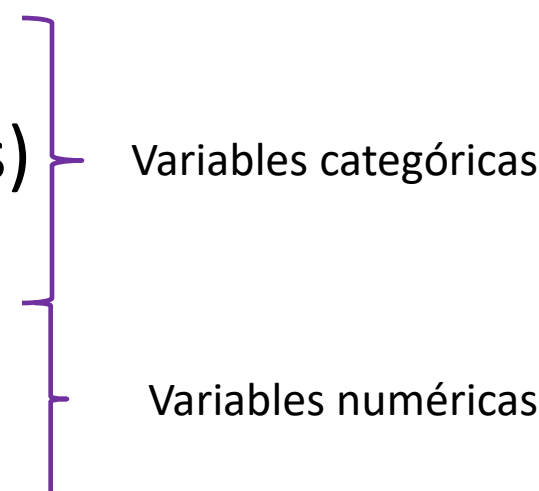
Interpretación gráfica



Conjunto de datos Salaries, de la librería de R “carData”

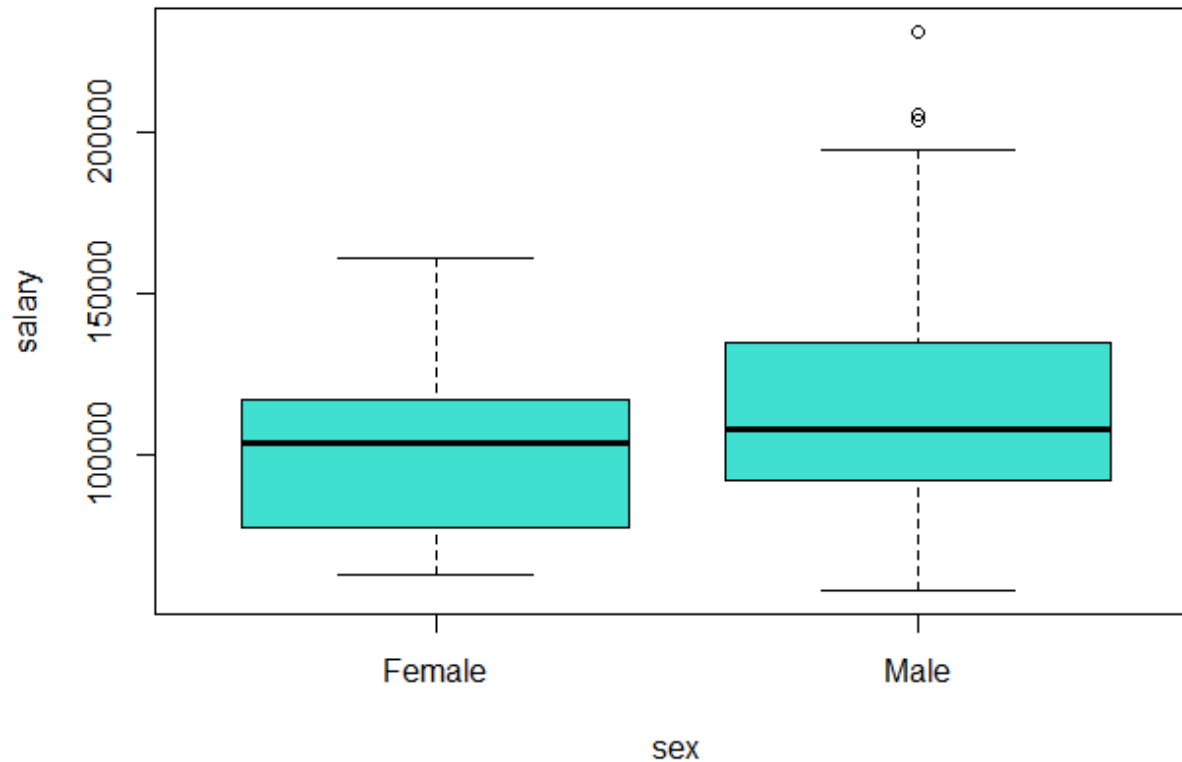
Tenemos 397 observaciones y 6 variables.

Queremos predecir el salario de profesores universitarios (salary) en función de:

- rank: AssitProf, AssocProf, Prof
 - discipline: A (teóricos), B (aplicados)
 - sex: M, F
 - yrs.since.phd
 - yrs.service
- 
- Variables categóricas
- Variables numéricas

Comenzamos haciendo la regresión utilizando como único regresor la variable sex.

```
boxplot(salary~sex,col="turquoise",data=Salaries)
```



Comenzamos haciendo la regresión utilizando como único regresor la variable sex.

```
model <- lm(salary ~ sex, data = Salaries)
```

```
summary(model)
```

Call:

```
lm(formula = salary ~ sex, data = Salaries)
```

Residuals:

Min	1Q	Median	3Q	Max
-57290	-23502	-6828	19710	116455

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	101002	4809	21.001	< 2e-16 ***
sexMale	14088	5065	2.782	0.00567 **

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

Residual standard error: 30030 on 395 degrees of freedom

Multiple R-squared: 0.01921, Adjusted R-squared: **0.01673**

F-statistic: 7.738 on 1 and 395 DF, p-value: **0.005667**

La regresión es significativa, sin embargo el R^2 es muy bajo

El modelo estimado queda

$$\text{Salary} = 101002 + 14088 \text{ sex}[\text{Hombre}]$$

El salario promedio de las mujeres es \$101002, mientras que el de los hombres es \$115090

`contrasts(Salaries$sex)` # código de la variable.

Male

Female 0

Male 1

¿Qué pasa si la recodificamos?

```
library(tidyverse)
```

```
mutate(sex = relevel(sex, ref = "Male"))#recodifica
```

```
contrasts(Salaries$sex) # como codifica R las variables.
```

Female

Male 0

Female 1

```
model <- lm(salary ~ sex, data = Salaries) #corremos la regresión
```

```
summary(model)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	115090.42	1587.378	72.503463	2.459122e-230
sexFemale	-14088.01	5064.579	-2.781674	5.667107e-03

El modelo queda

Salary=115090-14088 sex[Mujer]

Las conclusiones no varían

¿Qué hubiera pasado si hubiéramos hecho un test de dos muestras?

```
t.test(salary ~ sex, var.equal=TRUE, data = Salaries)
```

Two Sample t-test

data: salary by sex

t = 2.7817, df = 395, p-value = **0.005667** → Mismo p-valor que el test t

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

4131.107 24044.910

sample estimates:

mean in group Male mean in group Female

115090.4

101002.4

→ Mismas medias para los dos grupos

Observación: Si por ejemplo hubieramos utilizado otra codificación:

$$D = \begin{cases} 2 & \text{la persona esta desempleada} \\ 1 & \text{la persona esta empleada} \end{cases}$$

Los modelos se pueden plantear del mismo modo, se van a modificar las estimaciones de los parámetros β pero no el resto de los estadísticos estimados. El intercept no tendrá una interpretación inmediata, sin embargo, con un poco más de trabajo se llegan a las mismas conclusiones.

Si tuviéramos tres grupos entonces deberíamos definir dos variables D_1 y D_2 ,

$$D_1 = \begin{cases} 1 & \text{la observación pertenece al grupo A} \\ 0 & \text{la observación no pertenece al grupo A} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{la observación pertenece al grupo B} \\ 0 & \text{la observación no pertenece al grupo B} \end{cases}$$

Luego,

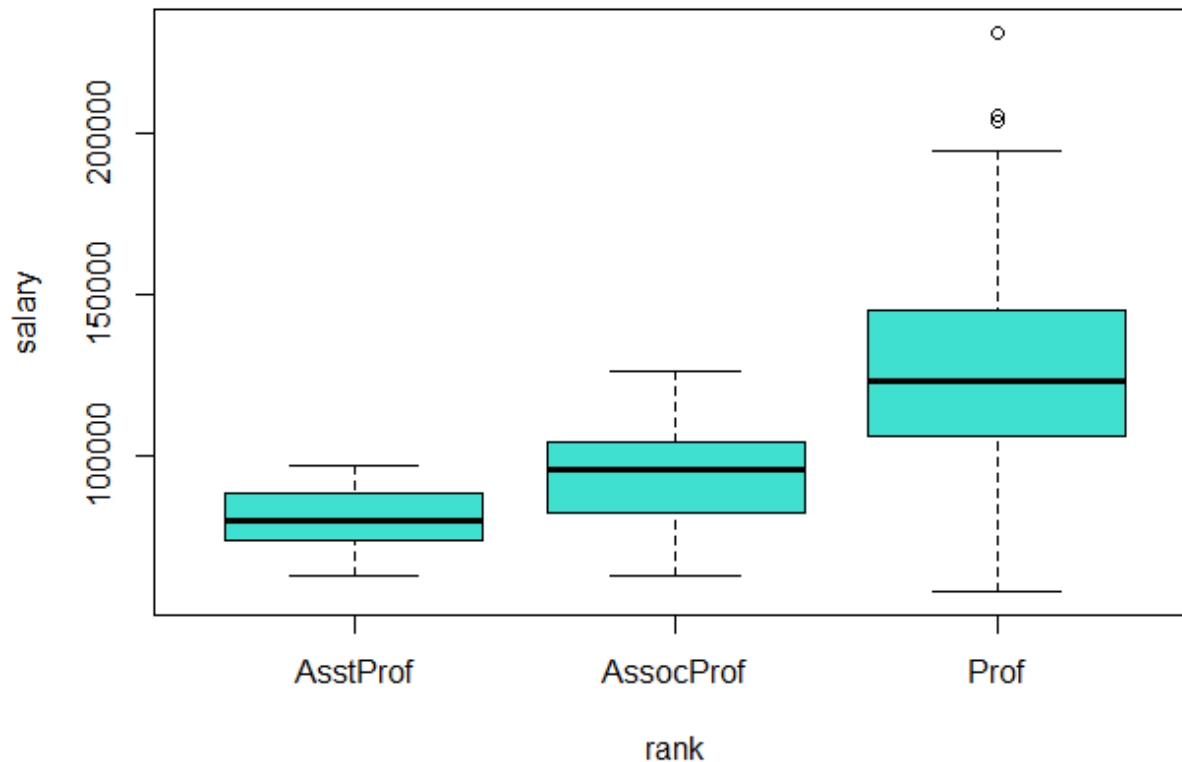
- ▶ Si $D_1 = 1$ y $D_2 = 0$ pertenece al grupo A.
- ▶ Si $D_1 = 0$ y $D_2 = 1$ pertenece al grupo B.
- ▶ Si $D_1 = 0$ y $D_2 = 0$ pertenece al grupo C.

el modelo de regresión en este caso sería

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2 + \epsilon.$$

Si la variable categórica tiene m categorías excluyentes hay que definir $m - 1$ variables dummy. Si se deciden usar m variables dummy entonces hay que descartar el intercept.

Corremos la regresión ahora utilizando la variable rank (cargo) que tiene tres categorías.
`boxplot(salary ~ rank, col="turquoise", data = Salaries)`



```
model <- lm(salary ~ rank, data = Salaries)
summary(model)
```

Call:

```
lm(formula = salary ~ rank, data = Salaries)
```

Residuals:

```
   Min    1Q  Median    3Q   Max
-68972 -16376 -1580  11755 104773
```

- Las categorías son significativas.
- La regresión es significativa.
- El R^2 aumentó.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	80776	2887	27.976	< 2e-16 ***
rankAssocProf	13100	4131	3.171	0.00164 **
rankProf	45996	3230	14.238	< 2e-16 ***

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

Residual standard error: 23630 on 394 degrees of freedom

Multiple R-squared: 0.3943, Adjusted R-squared: **0.3912**

F-statistic: 128.2 on 2 and 394 DF, p-value: **< 2.2e-16**

En este caso el modelo estimado queda

$$\text{salary} = 80776 + 13100 \text{ rank}[\text{Assoc}] + 45996 \text{ rank}[\text{Prof}]$$

- El salario promedio de un profesor asistente es \$80778
- El de un profesor asociado es \$93876.
- El de un profesor titular es \$126772.
- El coeficiente estimado 13100 es el efecto medio de pasar de ser un profesor asistente a asociado.
- El coeficiente estimado 45996 es el efecto medio de pasar de ser un profesor asistente a titular.

Análisis de la varianza de un factor

Supongamos que tenemos k muestras cada una de tamaño n cada una tiene una distribución $N(\mu_i, \sigma^2)$, $i = 1, \dots, k$. Es decir que las poblaciones difieren en sus **medias** pero no en sus varianzas. Es decir,

$$\begin{aligned} y_{ij} &= \mu_j + \epsilon_{ij} \text{ para } j = 1, \dots, n; i = 1, \dots, k. \\ &= \mu + (\mu_i - \mu) + \epsilon_{ij} \\ &= \mu + \tau_i + \epsilon_{ij}, \end{aligned}$$

donde

- ▶ y_{ij} es la j -ésima observación para el i -ésimo tratamiento fijo de efecto fijo $\tau_i = \mu_i - \mu$.
- ▶ μ es el efecto medio.
- ▶ ϵ_{ij} son los errores son iid, $\epsilon_{ij} \sim N(0, \sigma^2)$.

Como

$$\sum_{i=1}^k \tau_i = \sum_{i=1}^k (\mu_i - \mu) = \left(\sum_{i=1}^k \mu_i \right) - k\mu = 0$$

Luego,

$$H_0 : \tau_1 = \cdots = \tau_k = 0 \text{ vs } H_A : \exists i \text{ tal que } \tau_i \neq 0.$$

Buscamos los estimadores de μ y τ_i por mínimos cuadrados.

$$RSS = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \mu - \tau_i)^2$$

$$\frac{\partial RSS}{\partial \mu} = 0 \text{ tenemos } \hat{\mu} = \frac{1}{kn} \sum_{i=1}^k \sum_{j=1}^n y_{ij} = \bar{y}$$

$$\frac{\partial RSS}{\partial \tau_i} = 0 \text{ tenemos } \hat{\tau}_i = \frac{1}{n} \sum_{j=1}^n (y_{ij} - \hat{\mu}) = \bar{y}_i - \bar{y}$$

donde $\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij}$.

El test correspondiente es

$$F_0 = \left[\frac{\frac{n}{k-1} \sum_{i=1}^k (\bar{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}{k(n-1)}} \right] = \frac{S_{between}^2}{S_{within}^2}$$

donde $F_0 \sim f_{k-1, k(n-1)}$ bajo H_0 .

Rechazamos H_0 si $F_0 > f_{k-1, k(n-1), 1-\alpha}$, concluyendo en ese caso que los tratamientos no son idénticos.

Implementación en R, queremos predecir el salario en función del cargo.

#Hacemos el análisis de la varianza, el test F

```
model.aov <- aov(salary ~ rank, data = Salaries)
```

Summary of the analysis

```
summary(model.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rank	2	1.432e+11	7.162e+10	128.2	<2e-16 ***
Residuals	394	2.201e+11	5.586e+08		

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

Luego hay diferencia significativa entre los grupos, es decir entre los salarios en los diferentes cargos.

Esto ya lo habíamos visto al armar el modelo con los tres factores.

Comparaciones múltiples

Una vez que con el test F determinamos que la variable cualitativa es útil para predecir la variable de respuesta, es natural querer comparar las medias de los diferentes grupos.

Es decir comparar μ_i y μ_k , i.e. $H_0 : \mu_i - \mu_k = 0$.

Si las comparaciones se hacen de a pares hay que hacer $\frac{m(m-1)}{2}$ comparaciones, si cada test tiene nivel α la probabilidad global de cometer error tipo 1 es,

$$(1 - \alpha)^{\frac{m(m-1)}{2}}.$$

Si por ejemplo tenemos tres grupos como en el caso del salario de profesores predichos a partir de su cargo (rank), tenemos que el nivel es

$$(1 - 0.05)^{\frac{3(3-1)}{2}} = 0.95^3 = 0.8573.$$

Además esto afecta la potencia, los intervalos de confianza, etc.

Problemas de la salida por default R (y cualquier otro software):

- ▶ Los p-valores que da son para regresores particulares, en este caso son sobre una misma variable.
- ▶ No tenemos las comparaciones de a pares.

Luego, queremos testear **simultáneamente**

$$H_0 = \begin{cases} \mu_1 - \mu_2 & = & 0 \\ \mu_1 - \mu_3 & = & 0 \\ \mu_3 - \mu_2 & = & 0 \end{cases}$$

versus H_A : alguna de las igualdades no se verifica.

Con este análisis se busca determinar cuales son las categorías que tienen media diferente. Si encontramos dos grupos cuya media no difiere . se recomienda reagruparlos en uno ya que se estimaría peor la varianza, por otro lado también se dificultaría la interpretación del problema.

Método de Tukey

Es un procedimiento mediante el cual se pueden hacer todas las comparaciones de a pares con un nivel prefijado α .

```
compmult.tukey<-TukeyHSD(aov(salary ~ rank,data=Salaries),"rank")
compmult.tukey
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = salary ~ rank, data = Salaries)

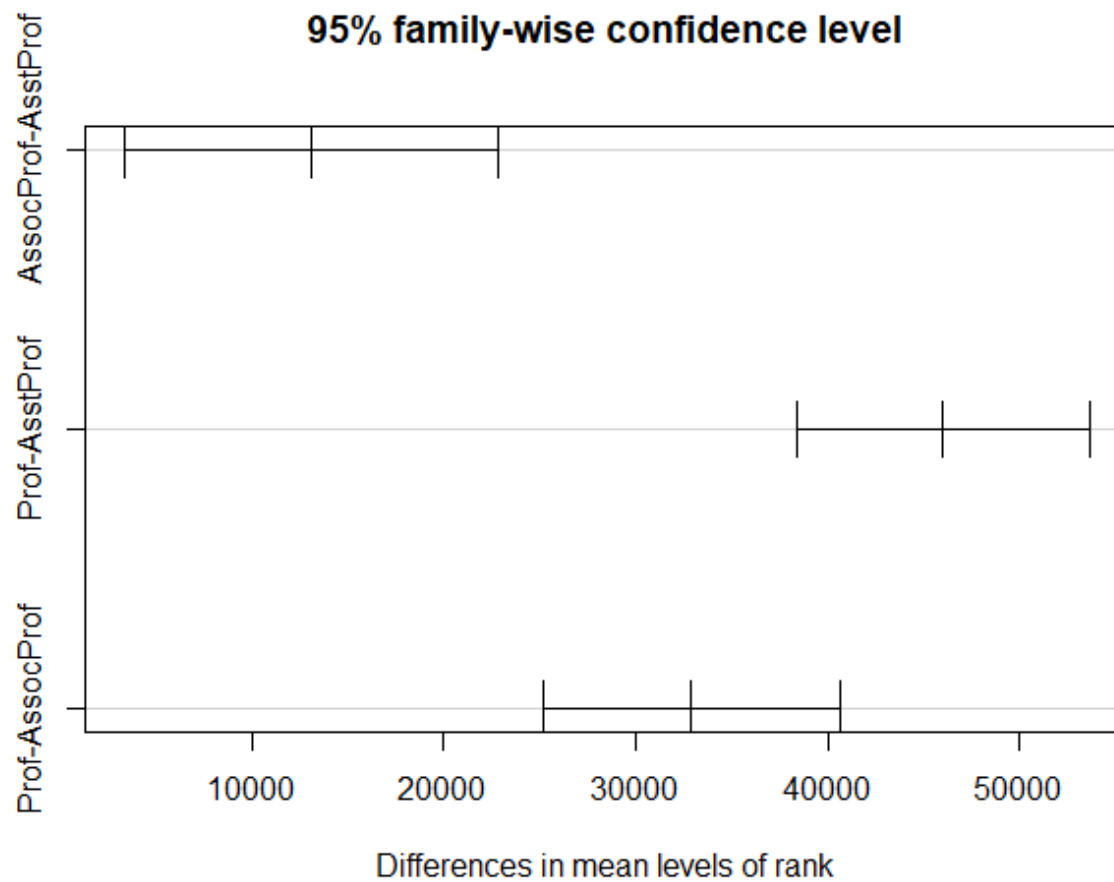
\$rank

	diff	lwr	upr	p adj
AssocProf-AsstProf	13100.45	3382.195	22818.71	0.0046514
Prof-AsstProf	45996.12	38395.941	53596.31	0.0000000
Prof-AssocProf	32895.67	25154.507	40636.84	0.0000000

Dif de medias	Lim Inf IC	Lim Sup IC	p-valor ajustado
------------------	---------------	---------------	---------------------

En este caso todas
las diferencias son
significativas.

`plot(compmult.tukey)`



Observación: puede resultar tentador la idea de usar una escala ordinal para una variable discreta, para no tener que utilizar $m - 1$ variables.

Ejemplo

$$Rank = \begin{cases} 1 & \text{Assistant Professor} \\ 2 & \text{Associate Professor} \\ 3 & \text{Full Professor} \end{cases}$$

Presenta dificultades:

- ▶ La escala es arbitraria.
- ▶ Se impone la misma distancia entre Assistant a Associate y de Associate a Full y esto puede no reflejar la realidad.
- ▶ En los casos donde la variable no presenta un orden imponer uno lleva a problemas de interpretación.

Interacciones

Consideremos ahora el caso en que tengamos dos variables regresoras

- ▶ x una variable continua,
- ▶ D una variable dummy

y que estas interactúan entre sí, luego consideramos el modelo

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 \underbrace{(D_i * x_i)}_{\text{interacción}} + \epsilon_i.$$

con $E(\epsilon_i) = 0$ y $\text{var}(\epsilon_i) = \sigma^2$

Luego,

$$E(y_i | D_i = 0) = \beta_0 + \beta_1 x_i$$

$$E(y_i | D_i = 1) = \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i$$

Tienen diferente pendiente e intercept

► β_2 representa el cambio de intercept cuando $D_i = 1$

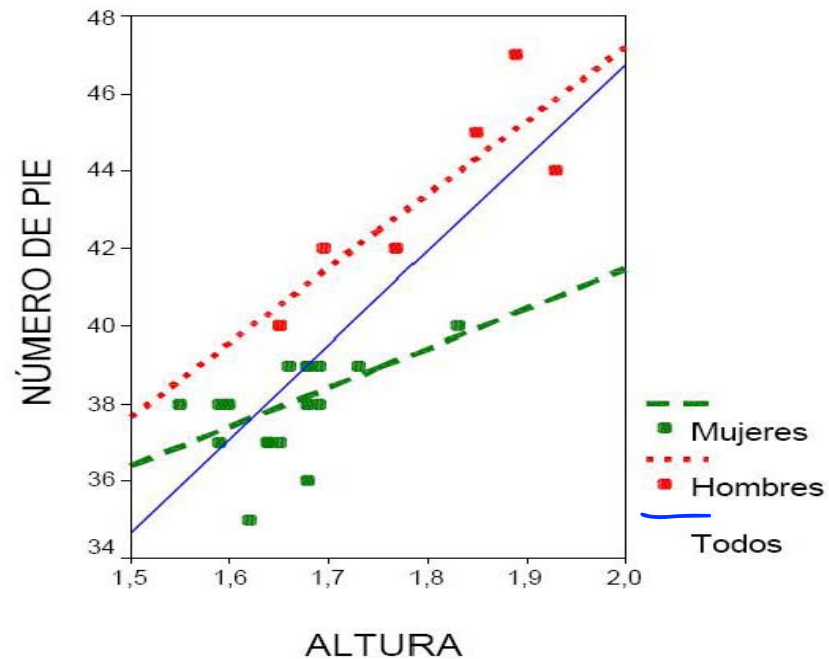
► β_3 representa el cambio de pendiente cuando $D_i = 1$

Ajustar el modelo con interacción es equivalente a ajustar por separado,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \epsilon_i.$$

Interpretación gráfica



En este caso se puede ver que cambia tanto la pendiente como el intercept.

Test de hipótesis vinculados

Si queremos testear si las dos regresiones son idénticas, entonces

$$H_0 : \beta_2 = \beta_3 = 0 \text{ vs } H_A : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0.$$

No rechazar H_0 indica que un solo modelo es necesario.

Si queremos testear que los dos modelos solamente difieren en el intercept y tienen la misma pendiente entonces

$$H_0 : \beta_3 = 0 \text{ vs } H_A : \beta_3 \neq 0.$$

Retomamos nuestro ejemplo sumando un regresor continuo en este caso *años de servicio* (yrs.service)

El modelo que queremos ajustar es

$$salary_i = \beta_0 + \beta_1 sex[Female]_i + \beta_2 yrs.service_i + \epsilon_i$$

para $i = 1, \dots, 397$ donde los errores siguen los supuestos usuales.

```
model <- lm(salary ~ sex+yrs.service, data = Salaries)
summary(model)
```

Call:

```
lm(formula = salary ~ sex + yrs.service, data = Salaries)
```

Residuals:

```
    Min     1Q  Median     3Q      Max
-81757 -20614 -3376  16779 101707
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	101428.7	2531.9	40.060	< 2e-16 ***
sexFemale	-9071.8	4861.6	-1.866	0.0628.
yrs.service	747.6	111.4	6.711	6.74e-11 ***

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

Residual standard error: 28490 on 394 degrees of freedom

Multiple R-squared: 0.1198, Adjusted R-squared: **0.1154**

F-statistic: 26.82 on 2 and 394 DF, p-value: **1.201e-11**

El modelo ajustado queda

$$\text{salary} = 101428.7 - 9071.8 \text{sex}[\text{Female}] + 747.6 \text{yrs.service}$$

Luego, el salario estimado para mujeres es

$$\text{salary} = 92356.9 + 747.6 \text{yrs.service}$$

y para hombres es

$$\text{salary} = 101428.7 + 747.6 \text{yrs.service}$$

#En un gráfico

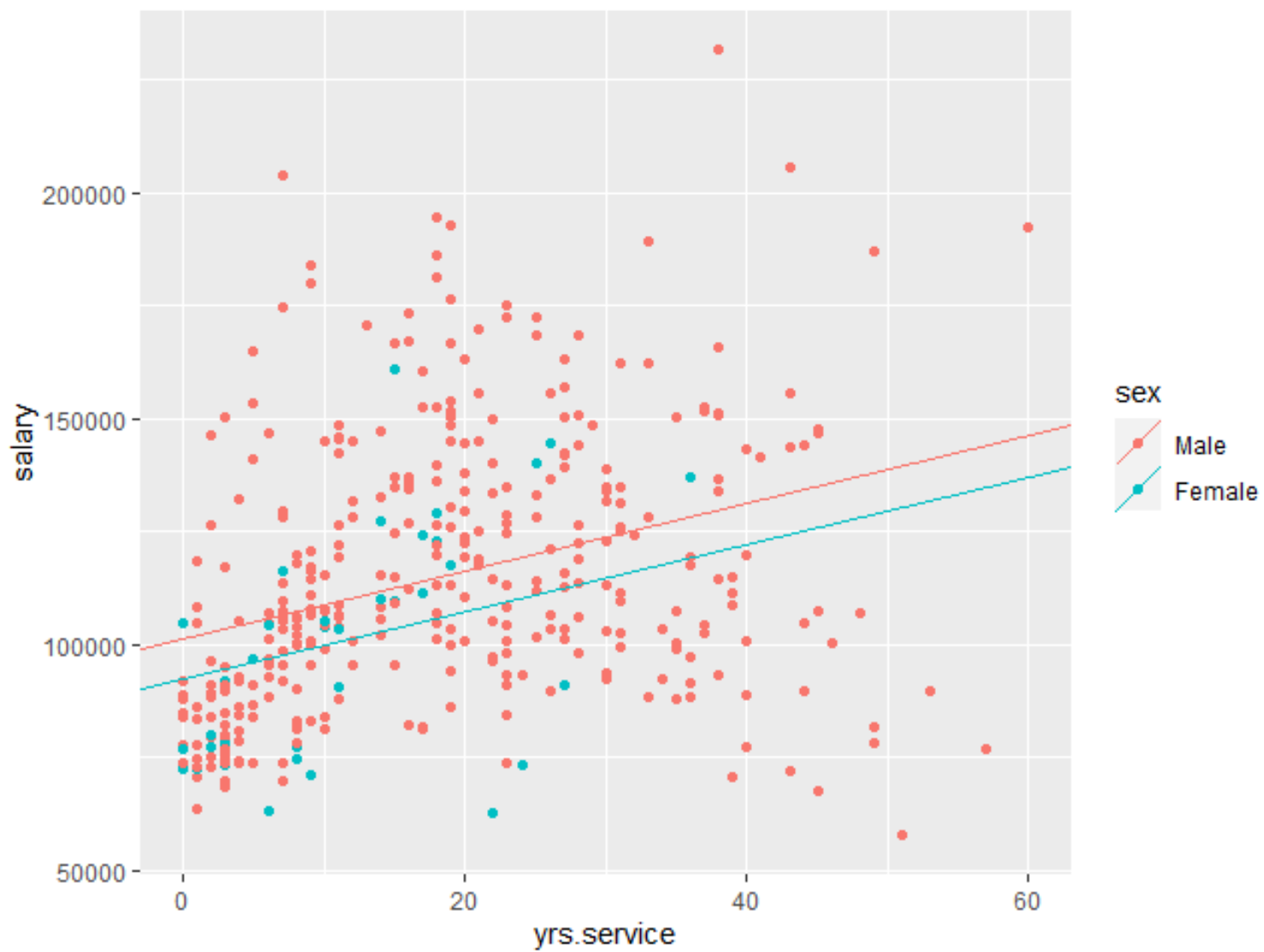
```
library(knitr)
```

Calcular los intercept específicos para la variable sex

```
intercepts <- c(coef(model)["(Intercept)"],  
                coef(model)["(Intercept)"] + coef(model)["sexFemale"])
```

```
lines.df <- data.frame(intercepts = intercepts,  
                      slopes = rep(coef(model)["yrs.service"], 2),  
                      sex = levels(Salaries$sex))
```

```
qplot(x = yrs.service, y = salary, color = sex, data = Salaries) +  
  geom_abline(aes(intercept = intercepts,  
                  slope = slopes,  
                  color = sex), data = lines.df)
```



Ajuste con una variable predictora cualitativa y otra numérica

Agregamos interacción entre las dos variables regresoras
El modelo que queremos ajustar es

$$\begin{aligned} salary_i = & \beta_0 + \beta_1 sex[Female]_i + \beta_2 yrs.service_i + \\ & + \beta_3 sex[Female]_i * yrs.service_i + \epsilon_i \end{aligned}$$

para $i = 1, \dots, 397$ donde los errores siguen los supuestos usuales.


```
model <- lm(salary ~ sex*yrs.service, data = Salaries)
summary(model)
```

Call:

```
lm(formula = salary ~ sex * yrs.service, data = Salaries)
```

Residuals:

```
    Min     1Q  Median     3Q      Max
-80381 -20258 -3727  16353 102536
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	102197.1	2563.7	39.863	< 2e-16 ***
sexFemale	-20128.6	7991.1	-2.519	0.0122 *
yrs.service	705.6	113.7	6.205	1.39e-09 ***
sexFemale:yrs.service	931.7	535.2	1.741	0.0825 .

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

Residual standard error: 28420 on 393 degrees of freedom

Multiple R-squared: 0.1266, Adjusted R-squared: **0.1199**

F-statistic: 18.98 on 3 and 393 DF, p-value: **1.622e-11**

El modelo ajustado queda

$$\text{salary} = 102197.1 - 20128.6 \text{ sex[Female]} + 705.6 \text{ yrs.service} + 931.7 \text{ sex[Female]} * \text{yrs.service}$$

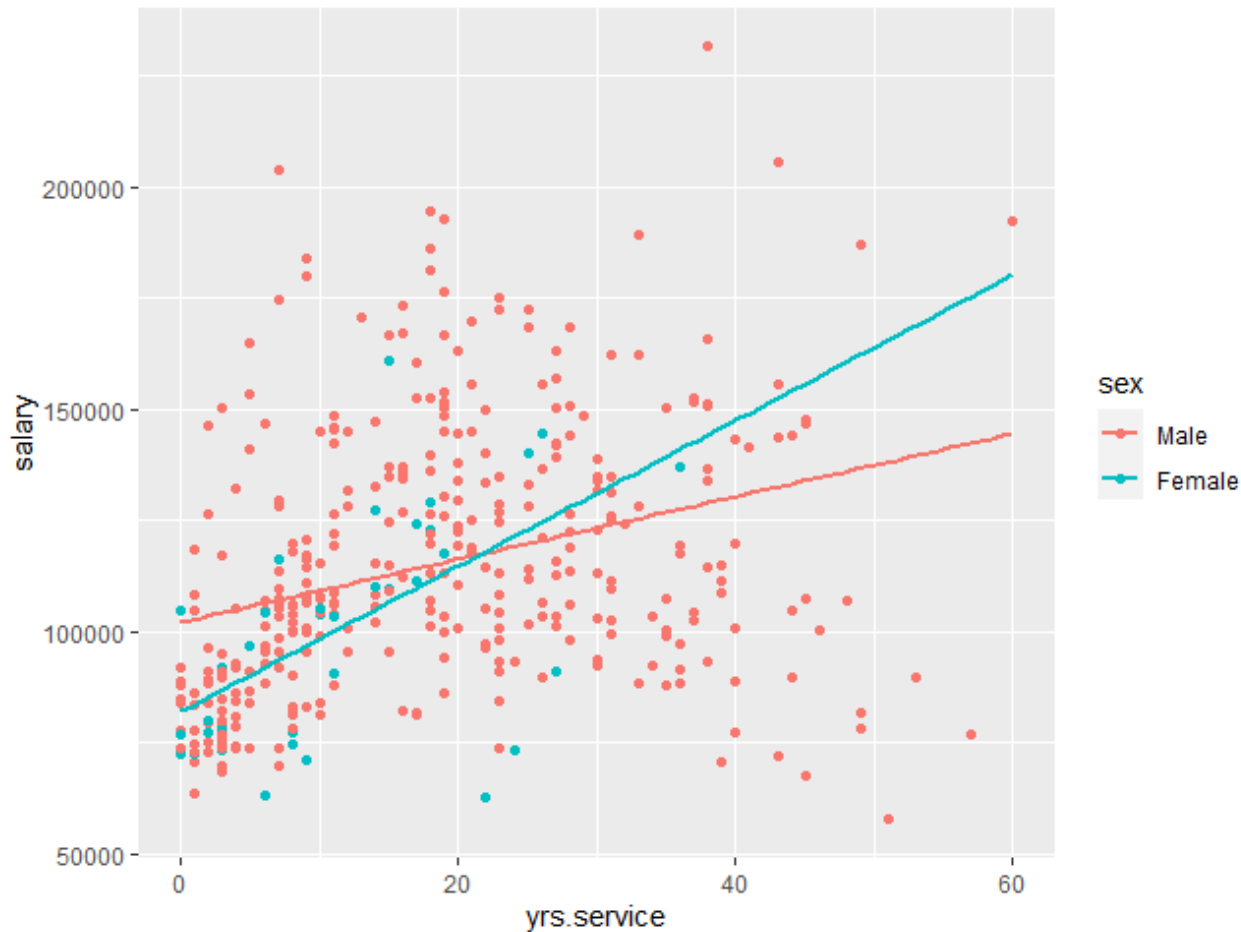
Luego, el salario estimado para mujeres es

$$\text{salary} = 82068.5 + 1637.3 \text{ yrs.service}$$

y para hombres es

$$\text{salary} = 102197.1 + 705.6 \text{ yrs.service}$$

```
qplot(x = yrs.service, y = salary, color = sex, data = Salaries) +  
  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



- Estos gráficos se pueden hacer a mano no hace falta usar esta fórmula.
- Vemos claro un problema de extrapolación, no hay mujeres que lleguen a los 40 años de servicio.

Variables indicadoras versus variables cuantitativas

En ocasiones variables cuantitativas como la edad, el ingreso, etc son difíciles de tomar con precisión.

Usualmente se las segmenta en categorías m que se pueden modelar con $m - 1$ variables indicadoras.

Desventajas:

- ▶ Aumenta la cantidad de parámetros del modelo.
- ▶ No se trabaja con información precisa.

Ventajas:

- ▶ Se puede utilizar esta información aunque no se tenga con precisión.

Si hay muchos datos la disminución de los grados de libertad no es un problema en caso contrario sí.

Interacción entre dos variables cuantitativas

Retomemos el ejemplo Advertising, habíamos concluido que la publicidad en TV y radio tiene efecto en las ventas.

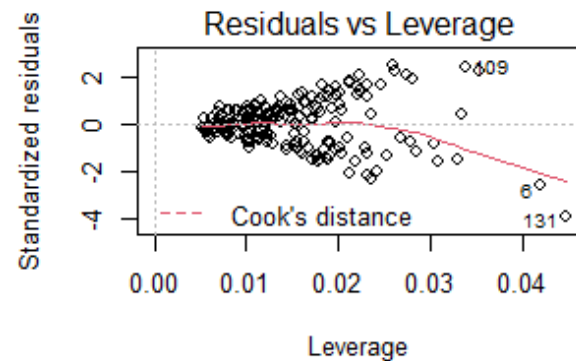
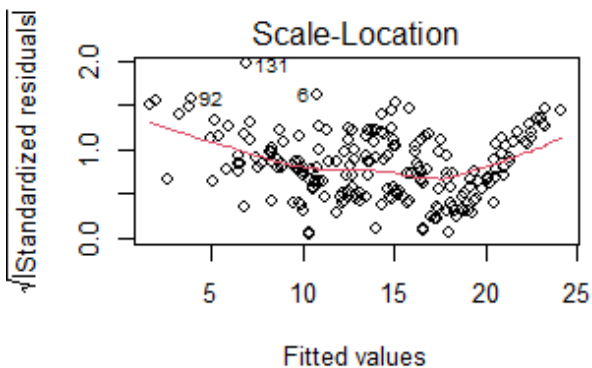
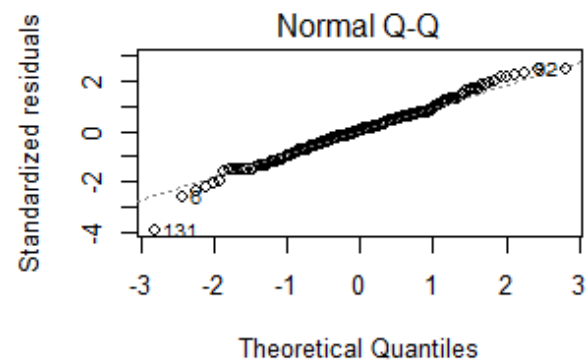
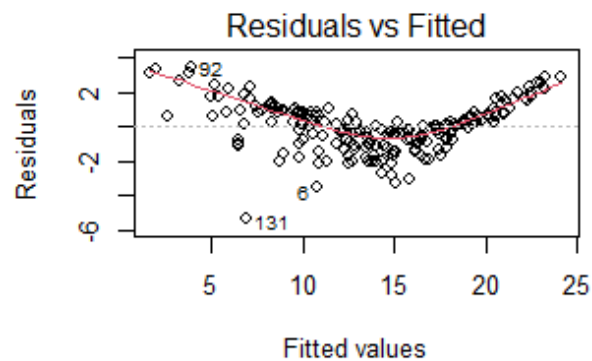
Con el modelo que construimos estas publicidades tienen efectos independientes y no sinergizan.

El modelo establece una tasa constante de aumento en las ventas por cada mil pesos invertido en publicidad que está dado por el β_1 y en forma análoga para la radio. Recordemos que el modelo que teníamos era

$$sales_i = \beta_0 + \beta_1 \sqrt{TV_i} + \beta_2 radio_i + \epsilon_i,$$

para $i = 1, \dots, 200$.

Recordemos que estos son los gráficos de los residuos del modelo

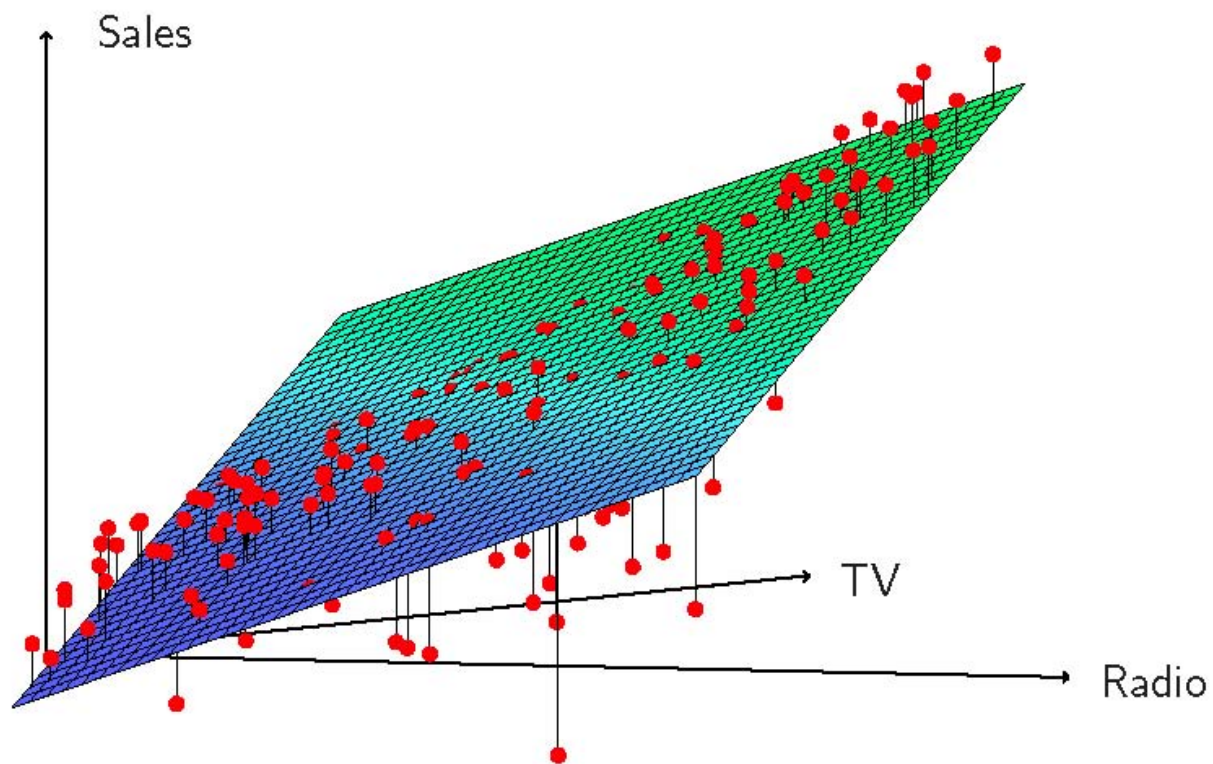


Al analizar los residuos vemos que el modelo parece sobreestimar las ventas para determinadas regiones y subestimarlas para otras. Si ajustamos un modelo más sencillo

$$sales_i = \beta_0 + \beta_1 \sqrt{TV_i} + \beta_2 radio_i + \epsilon_i,$$

para $i = 1, \dots, 200$, vemos que el modelo subestima las ventas de un solo medio y sobreestima las de dos medios.

La presencia de no linealidad en los residuos es un posible indicio de que puede haber sinergia entre las variables.



El supuesto de **aditividad** podría ser erróneo.

Qué pasaría si invertir plata en publicidad en radio hiciera que la publicidad en TV fuese más efectiva?

La pendiente de la TV debería aumentar al aumenta cuando lo hace la de la radio. Es decir si podemos realizar una inversión de \$10000, sería conveniente poner mitad en cada medio y no todo en un medio.

Consideramos el siguiente modelo

$$sales_i = \beta_0 + \beta_1 \sqrt{TV_i} + \beta_2 radio_i + \beta_3 \sqrt{TV_i} * radio_i + \epsilon_i,$$

```
adv.lm.4=lm(sales~sqTV*radio)
```

```
summary(adv.lm.4)
```

Call:

```
lm(formula = sales ~ sqTV * radio)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.0562	-0.2757	-0.0121	0.2758	1.2421

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.4444112	0.1793714	24.778	< 2e-16 ***
sqTV	0.4383960	0.0150223	29.183	< 2e-16 ***
radio	-0.0500957	0.0062645	-7.997	1.09e-13 ***
sqTV:radio	0.0215106	0.0005179	41.538	< 2e-16 ***

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

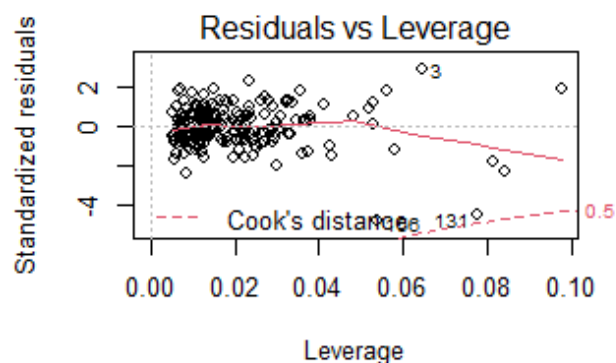
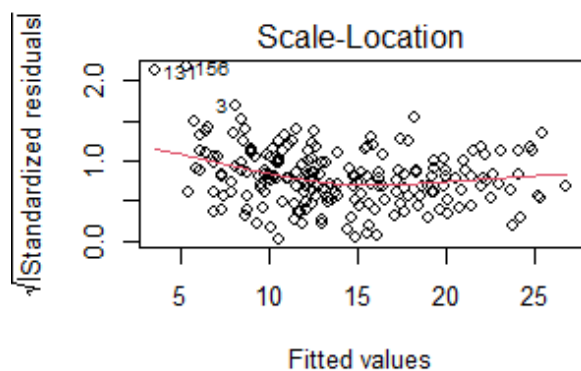
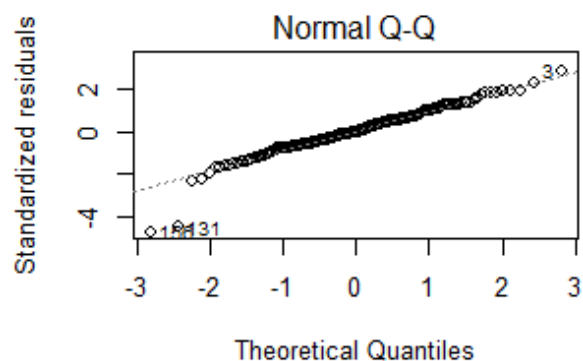
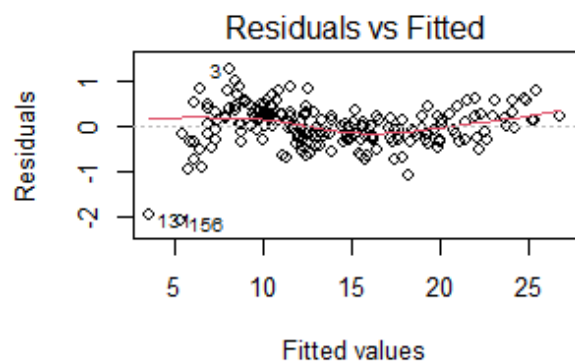
Residual standard error: 0.4476 on 196 degrees of freedom

Multiple R-squared: 0.9928, Adjusted R-squared: 0.9926

F-statistic: 8949 on 3 and 196 DF, p-value: < 2.2e-16

El modelo ajustado es

$$\widehat{sales} = 4.44 + 0.438\sqrt{TV} - 0.05radio + .021\sqrt{TV} * radio$$



- Interpretación de las interacciones y su efecto sobre la linealidad del modelo.
- Ejemplo:
 - Se quiere estudiar la productividad de una empresa. Se quiere predecir la cantidad de unidades que se fabricaran (Y), conociendo la cantidad de líneas de producción (X_1) y la cantidad de empleados (X_2). Es claro que el efecto de agregar una línea de producción va a estar vinculado al número de trabajadores disponibles para operar en las líneas.

Regresión lineal: Removiendo el problema de Aditividad.

unidades

$$\approx 1.2 + 3.4 \times \text{líneas} + 0.22$$

$$\times \text{trabajadores} + 1.4 \times \text{líneas}$$

$$\times \text{trabajadores} =$$

$$1.2 + (3.4 + 1.4 \times \text{trabajadores}) \times \text{líneas} \\ + 0.22 \times \text{trabajadores}.$$

Agregar una línea de producción va a incrementar la producción en $(3.4 + 1.4 \times \text{trabajadores})$ unidades.

Interpretación en nuestro ejemplo Podemos reescribir el modelo del siguiente modo

$$sales_i = \beta_0 + (\beta_1 + \beta_3 radio) \sqrt{TV}_i + \beta_2 radio_i + \epsilon_i,$$

- ▶ β_3 es el incremento en la efectividad de las ventas en \sqrt{TV} por cada unidad que se incrementan LA INVERSIÓN en radio o viceversa.
- ▶ Por cada \$1000² que se inviertan en TV, las ventas se van a ver incrementadas en $438 + 21 \times radio$ unidades. En ocasiones ocurre que los p-valores de los efectos principales (las variables originales, en nuestro caso \sqrt{TV} y Radio) son grandes. En estos casos la jerarquía indica que si se incluye el efecto de la interacción los efectos principales deben estar incluidos en el modelo.

Bibliografía

- Quinn, G, Keough, M. (2002). Experimental Design and Data Analysis for Biologists, Cambridge University Press. Capítulos 7 y 8.
- Harrell, F. (2015), Regression Modelling Strategies, 2nd Edition , Ed Springer. Capítulo 2.