Regresión Lineal en R

Sesión 3

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Datos

La base de datos `Happyness` contiene información sobre la percepción de satisfacción de las personas con su vida. Se les pide a los encuestados que piensen en una escalera con la mejor vida posible para ellos con un 10 y la peor vida posible con un 0 (escalera de Cantril) y que califiquen sus propias vidas actuales en esa escala.

Las columnas que siguen al puntaje de felicidad estiman la medida en que cada uno de los seis factores (producción económica, apoyo social, esperanza de vida, libertad, ausencia de corrupción y generosidad) contribuyen a que las evaluaciones de vida sean más altas en cada país.

Interesa ajustar una regresión lineal para modelar los puntajes de percepción de satisfacción en término de las demás covariables.

Contexto

3

Un equipo de sociología plantea que la producción económica (`GDP per capita`) y los puntajes promedios de percepción de satisfacción (`Score`) se asocian positivamente y además, se observa una relación lineal fuerte y bastante marcada. Interesa corroborar esta teoría.

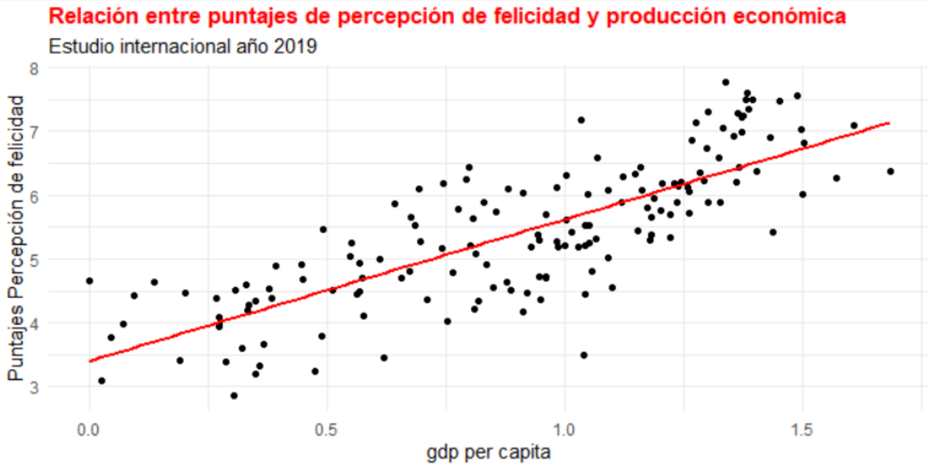
Análisis previo

```
puntaje=Happyness$Score   #Variable respuesta

gdp=Happyness$`GDP per capita` #Variable explicativa

summary(puntaje)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.853   4.545   5.380   5.407   6.184   7.769
summary(gdp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.6028  0.9600  0.9051  1.2325  1.6840

cov(puntaje,gdp)   #Asociacion positiva
[1] 0.3520515

cor(puntaje,gdp)   #Correlacion lineal es alta
[1] 0.7938829

cor(puntaje,gdp,method="spearman")
[1] 0.8144834

#Correlacion de pearson y spearman son muy similares
```

Gráfico de dispersión

```
#Grafico de dispersion

#install.packages("ggplot2")

library(ggplot2)

df<-data.frame(puntaje,gdp)

graph<-ggplot(df, aes(x = gdp, y = puntaje))+geom_point()+
  ggtitle("Relacion entre puntajes de percepcion de felicidad y
  produccion economica")+xlab("gdp per capita")+
  ylab("Puntajes Percepcion de felicidad")+
  labs(subtitle="Estudio internacional año 2019")+theme_minimal()+
  theme(
    plot.title = element_text(color = "red", size = 13, face = "bold"))


graph+geom_smooth(method='lm', formula= y~x,col="red",se = FALSE)
```

Gráfico


Relación entre puntajes de percepción de felicidad y producción económica
Estudio internacional año 2019
Puntajes Percepción de felicidad
8
7
6
5
4
3
0.0
0.5
1.0
1.5
gdp per capita

Regresión Lineal simple

El modelo de regresión lineal simple con intercepto, es:

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

Donde ϵ_i se asume que distribuye normal $N(0, \sigma^2)$ y que los pares ϵ_i y ϵ_j con $i \neq j$ son independientes.

Así, los supuestos de un modelo de regresión lineal son:

- Normalidad de ϵ_i
- Homocedasticidad de los errores
- Independencia entre ϵ_i y ϵ_j con $i \neq j$
- Linealidad de la media de y, es decir, la media de y puede expresarse como una combinación lineal de los predictores

Regresión lineal en R

```
linealsimple<-lm(puntaje~gdp)

linealsimple

Call:
lm(formula = puntaje ~ gdp)

Coefficients:
(Intercept)          gdp
      3.399        2.218

coef(linealsimple)   #Coeficientes estimados

(Intercept)         gdp
   3.399345    2.218148
```

Interpretación

El modelo obtenido es:

$$\widehat{\text{puntaje}}_i = 3.399 + 2.218\text{gdp}_i$$

Considerando los errores se tiene:

$$\text{puntaje}_i = 3.399 + 2.218\text{gdp}_i + \hat{\epsilon}_i$$

- El intercepto es 3.399, es decir, cuando gdp es nulo se espera un puntaje alrededor de 3.399
- La pendiente es 2.218, es decir, a medida que gdp aumenta en una unidad, el puntaje de felicidad estimado/medio aumentaría en 2.218 unidades

Resumen del modelo

```
summary(linealsimple)

Call:
lm(formula = puntaje ~ gdp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22044 -0.48361  0.00828  0.48433  1.47409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3993     0.1353   25.12   <2e-16 ***
gdp           2.2181     0.1369   16.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.679 on 154 degrees of freedom
Multiple R-squared:  0.6303,Adjusted R-squared:  0.6278
F-statistic: 262.5 on 1 and 154 DF,  p-value: < 2.2e-16
```

Test T de significancia

Test t de significancia para los coeficientes contrasta las siguientes hipótesis

$$\mathrm{H}_0 : \beta_j = 0 \qquad \mathrm{H}_1 : \beta_j \neq 0$$

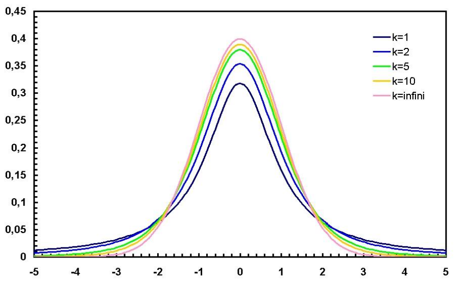
El test se construye:

$$t_j = \frac{\hat{\beta}_j}{s_{\beta_j}}$$

t_j tiene distribución T Student con parámetro $n - k - 1$ donde n es la cantidad de observaciones y k es la cantidad de variables.

Grados de libertad

Los grados de libertad corresponden a cuántos grados de información libre tenemos disponible. Por lo general, los grados de libertad son iguales al tamaño de la muestra menos la cantidad de parámetros que se necesita calcular durante un análisis. Mientras más parámetros utilicemos, menor es la cantidad de grados de libertad pues para calcular cada parámetro se requieren recursos (observaciones).



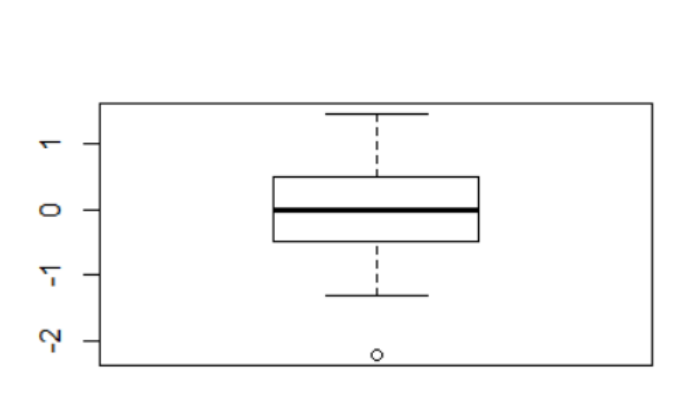
Ejemplo

Testeemos la significancia de β_1:

```
#Testeemos b1 (la pendiente)

estimacionbeta1<-coef(linealsimple)[2]

desviacionbeta1<-sqrt(vcov(linealsimple)[2,2])

tbeta1<-estimacionbeta1/desviacionbeta1

n<-nrow(Happyness)
k<-1

pt(-abs(tbeta1),df=(n-k-1)) #valor p menor que 0.05
2.15774e-35

# se rechaza que beta1 sea cero
```

Realizamos un boxplot de los residuos del modelo y observamos que existía una observación atípica:

```
boxplot(residuals(linealsimple))
```



Identificando outliers

```
#Identificando cual o cuales son outlier:

outliers<-which(residuals(linealsimple) < boxplot(residuals(linealsimple),plot=FALSE)$stats[1]|
residuals(linealsimple) > boxplot(residuals(linealsimple),plot=FALSE)$stats[5])

outliers
148
148

#Cual es el o los valores de los residuos atipicos?

residuals(linealsimple)[outliers]
      148
-2.220437
```

R^2

```
summary(linealsimple)$r.squared
[1] 0.63025
```

Mide la proporcion de varianza de la variable dependiente explicada por la variable dependiente

Solo utilizando la variable gdp se logra explicar un 63% de la variabilidad de los puntajes de percepción de felicidad

Cuando se tiene más de una variable se usa el R2 ajustado

```
summary(linealsimple)$adj.r.squared  #Ajusta el R2 por cantidad de
variables

[1] 0.627849
```

AIC y BIC

Son medidas para comparar el desempeño entre distintos modelos. Mientras menor es mejor. El BIC penaliza por cantidad de variables.

```
AIC(linealsimple)
[1] 325.9328

BIC(linealsimple)
[1] 335.0824
```

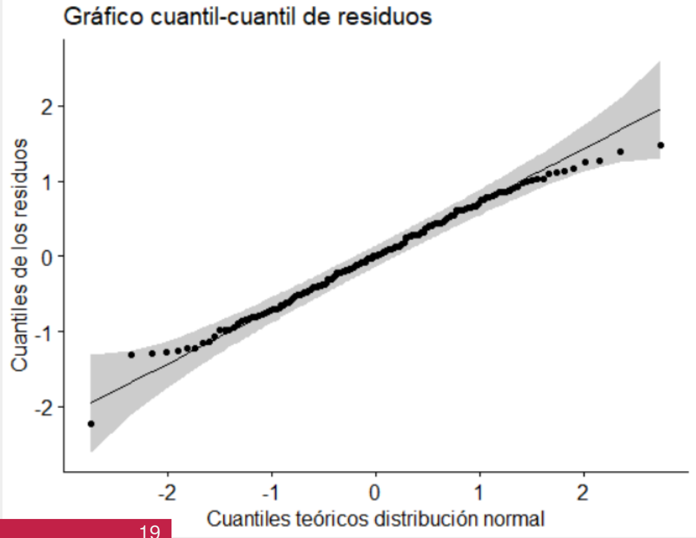
```
confint(linealsimple) #Intervalo de confianza de betas
               2.5 %   97.5 %
(Intercept) 3.132016 3.666675
gdp         1.947689 2.488607
```

```
vcov(linealsimple)  #Matriz de varianzas-covarianzas
            (Intercept)         gdp
(Intercept)  0.01831240 -0.01696582
gdp         -0.01696582  0.0187437
```

Residuos y valores ajustados

```
residuals(linealsimple)  #errores estimados

fitted(linealsimple)  #Estimaciones de los puntajes
```

Normalidad de ϵ_i



Test de normalidad

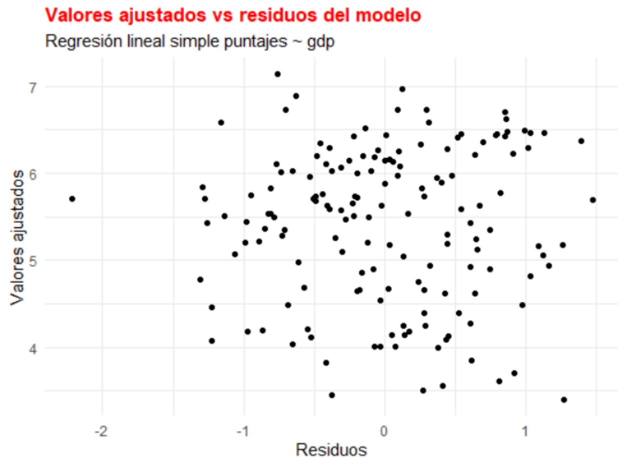
En el test de Shapiro-Wilk, se contrastan las siguientes hipótesis:

H_0 : Normalidad de ϵ
H_1 : No normalidad de ϵ

```
shapiro.test(residuals(linealsimple))

Shapiro-Wilk normality test

data:  residuals(linealsimple)
W = 0.99092, p-value = 0.4201
```

Homocedasticidad de ϵ_i

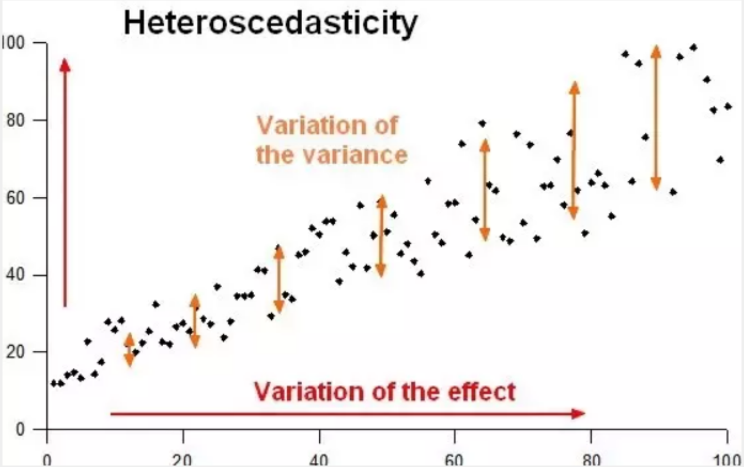
Se realiza un gráfico de dispersión entre los residuos y los valores ajustados del modelo:

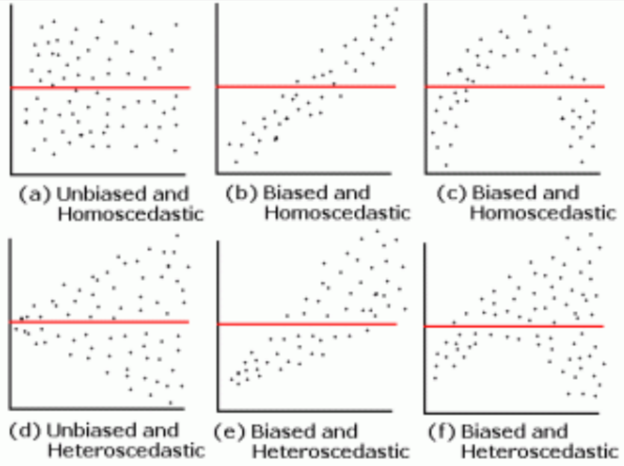


Valores ajustados y residuos

En el gráfico anterior, lo ideal es observar una nube de puntos sin ningún patrón o forma extraña.

Heterocedasticidad




(a) Unbiased and Homoscedastic
(b) Biased and Homoscedastic
(c) Biased and Homoscedastic
(d) Unbiased and Heteroscedastic
(e) Biased and Heteroscedastic
(f) Biased and Heteroscedastic

Test de Breusch-Pagan

Una vez que ya se ha realizado análisis de la significancia de las variables, se estudian los residuos del modelo y se realiza el test de homogeneidad de varianza de Breusch-Pagan. La hipótesis nula del test afirma que los residuos son homocedásticos y la hipótesis alternativa que los residuos son heterocedásticos.

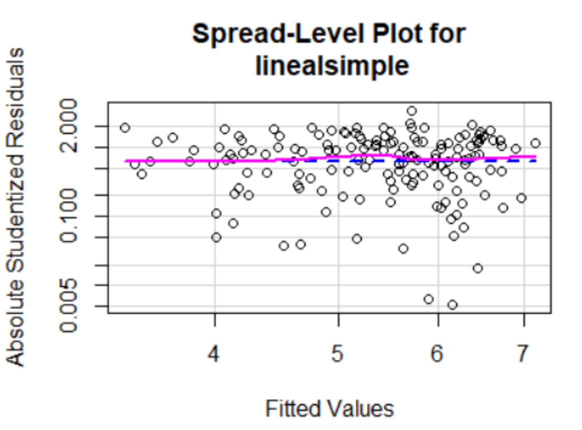
```
library(lmtest)
bptest(linealsimple)

studentized Breusch-Pagan test

data:  linealsimple
BP = 0.054916, df = 1, p-value = 0.8147

#Utilizando un 95% de confianza, no se rechaza la hipotesis de
homocedasticidad
```

Gráfico

Es posible crear un gráfico que permite analizar si existen cambios o no en la varianza dependiendo de las observaciones.

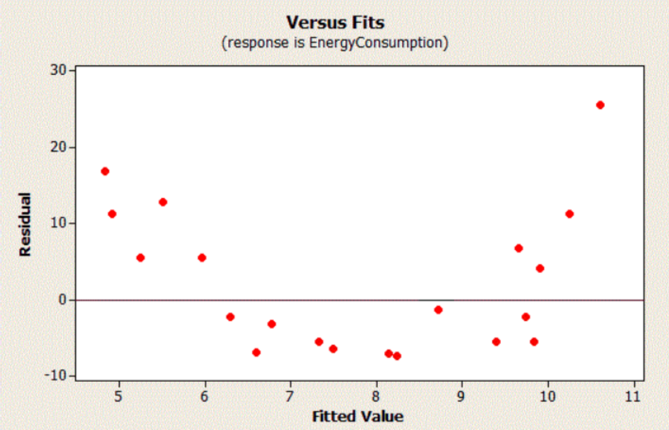
Gráfico en R

```
library(car)
spreadLevelPlot(linealsimple)
```



Mientras menos varíe la línea más homocedasticidad se observa.

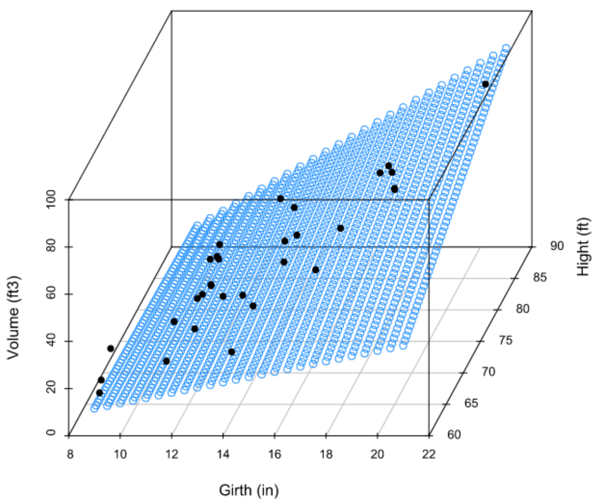
Es posible que exista algún nivel de correlación entre los residuos, pudiera ser provocado por algún efecto temporal.



Regresión lineal múltiple

Cuando se realiza un modelo de regresión con dos variables predictoras, ya no es posible realizar un gráfico de dispersión tan directo como anteriormente, pues ahora contamos con más dimensiones:

Dos variables explicativas



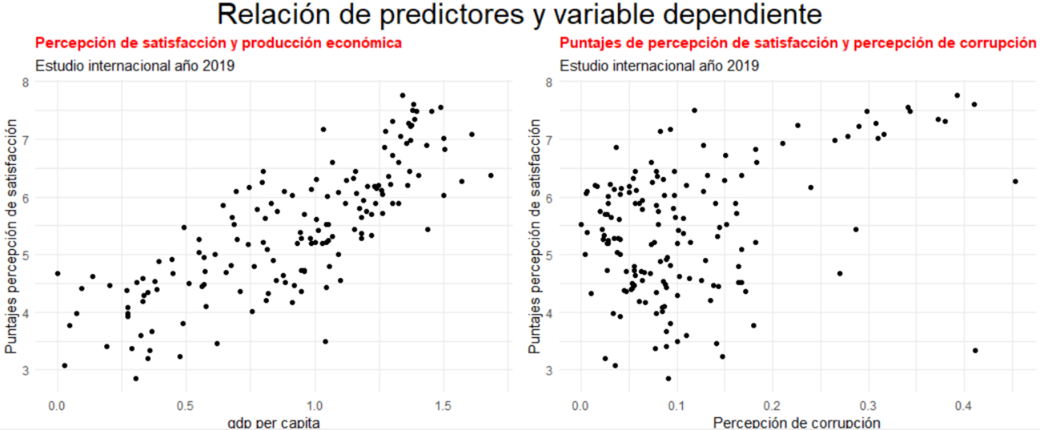
Contexto

Si bien, utilizando sólo la variable productividad económica para explicar los puntajes de percepción de satisfacción promedio se obtienen buenos resultados, otra variable de interés para incorporar en el modelo es percepción de corrupción y ver cómo está incide en los puntajes de percepción de satisfacción.

Análisis previo

```
corruption<-Happyness$‘Perceptions of corruption‘

cor(puntaje,gdp)
[1] 0.7938829

cor(puntaje,corruption)
[1] 0.3856131

cor(gdp,corruption)
[1] 0.2989198
```

Gráficos



Relación de predictores y variable dependiente
Percepción de satisfacción y producción económica
Estudio internacional año 2019
Puntajes percepción de satisfacción
8
7
6
5
4
3
0.0
0.5
1.0
1.5
gdp per capita
Puntajes de percepción de satisfacción y percepción de corrupción
Estudio internacional año 2019
Puntajes percepción de satisfacción
8
7
6
5
4
3
0.0
0.1
0.2
0.3
0.4
Percepción de corrupción

Gráfico 3D

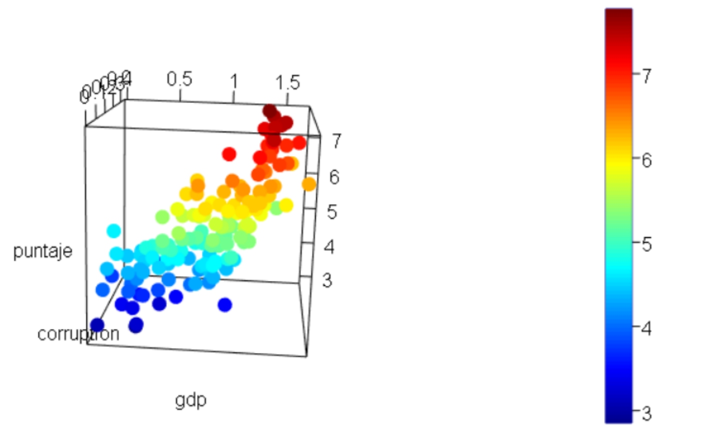
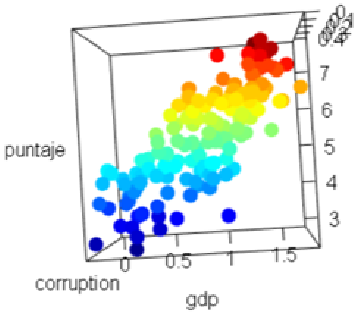


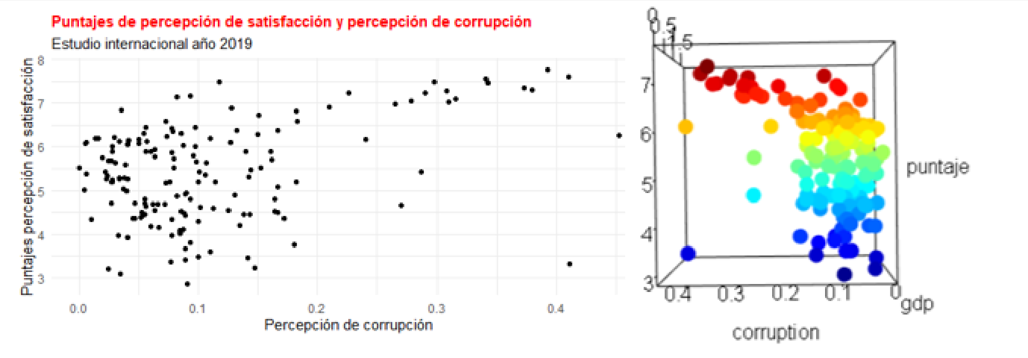
Gráfico 3D

Percepción de satisfacción y producción económica
Estudio internacional año 2019
Puntajes percepción de satisfacción
gdp per capita



puntaje
corruption
gdp

Gráfico 3D



Puntajes de percepción de satisfacción y percepción de corrupción
Estudio internacional año 2019
Puntajes percepción de satisfacción
Percepción de corrupción
0.0
0.1
0.2
0.3
0.4
3
4
5
6
7
8
puntaje
gdp
corruption

Modelo aditivo

```
multiple<-lm(puntaje~gdp+corruption)

coef(multiple)
(Intercept)         gdp  corruption
   3.310377    2.082130    1.917539
```

Modelo obtenido

- El intercepto estimado ($\hat{\beta}_0$) es 3.31, es decir, cuando gdp y corruption son nulos, se estima un puntaje de percepción de felicidad promedio de 3.31
- La magnitud asociada al desarrollo económico (gdp: $\hat{\beta}_1$)) es de 2.08, es decir, cuando corruption se mantiene constante y el desarrollo económico aumenta en una unidad, se observa que el puntaje de percepción de felicidad promedio aumenta en 2.08 unidades
- La magnitud asociada a la percepción de corrupción (corruption $\hat{\beta}_2$)) es de 1.91, es decir, cuando gdp se mantiene constante y corruption aumenta en una unidad, se observa que el puntaje de percepción de felicidad promedio aumenta en 1.91 unidades

Significancia de β

```
summary(multiple)

Call:
lm(formula = puntaje ~ gdp + corruption)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.18163 -0.45905  0.07395  0.45983  1.52537 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.3104     0.1340   24.70  < 2e-16 ***
gdp           2.0821     0.1392   14.96  < 2e-16 ***
corruption    1.9175     0.5864    3.27  0.00133 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6586 on 153 degrees of freedom
Multiple R-squared:  0.6544,Adjusted R-squared:  0.6499 
F-statistic: 144.9 on 2 and 153 DF,  p-value: < 2.2e-16
```

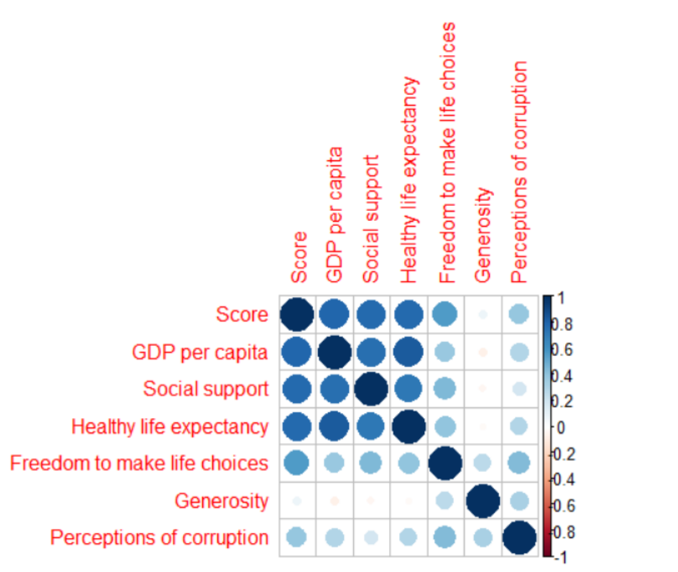
Tabla comparativa entre modelos

Modelo	p	R^2	AIC	BIC
`puntajes ~ gdp`	2	0.62	325.9	335
`puntajes ~ gdp+corruption`	3	0.65	317.3	329.5
`puntajes ~ gdp+freedom`	3	0.70	288.5	300.7
`puntajes ~ todas`	6	0.77	255.5	278.9

Correlación

```
#install.packages("corrplot")
library(corrplot)

cor<-cor(Happyness[,-c(1,2)])
corrplot(cor)
```

Correlación



Método Forward

Se van añadiendo variables de acuerdo algún criterio establecido. Se parte de un modelo sencillo y se van añadiendo más y más variables hasta que ya no se observe una mejora significativa al añadir más variables.

Método Forward

```
#Forward

library(MASS)

biggest<-formula(lm(Score~.,data=Happyness[,-c(1,2)]))
fwd.model = step(lm(Score ~ 1, data=Happyness[,-c(1,2)]),
direction='forward', scope=biggest)

fwd.model$call
lm(formula = Score ~ `GDP per capita` + `Freedom to make life choices` +
    `Social support` + `Healthy life expectancy` +
    `Perceptions of corruption`, data = Happyness[, -c(1,
    2)])

#En base al AIC, el orden de inclusion es:
# - GDP per capita
# - Freedom to make life choices
# - Social Support
# - Healthy life expectancy
# - Perceptions of corruption

#La variable Generosity no se incluye
```

Modelo obtenido por forward

```
summary(fwd.model)  #Todas son significativas

Residuals:
     Min       1Q   Median       3Q      Max
-1.82997 -0.35344  0.05803  0.35977  1.17522

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.8689     0.1973   9.471  < 2e-16 ***
‘GDP per capita‘                0.7455     0.2161   3.450 0.000728 ***
‘Freedom to make life choices‘  1.5340     0.3666   4.185 4.84e-05 ***
‘Social support‘                1.1180     0.2368   4.722 5.33e-06 ***
‘Healthy life expectancy‘       1.0840     0.3344   3.241 0.001467 **
‘Perceptions of corruption‘     1.1176     0.5218   2.142 0.033839 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 150 degrees of freedom
Multiple R-squared:  0.7777,Adjusted R-squared:  0.7703
F-statistic:   105 on 5 and 150 DF,  p-value: < 2.2e-16
```

Método Backward

Esta metodología va quitando variables dependiendo de qué tan costoso sea quitarlas. Parte con el modelo completo y a partir de ahí comienza a quitar aquéllas que significan menos costos para el modelo, es decir, quita las que aportan menos.

Método Backward

```
bcw.model<-step(lm(Score~.,data=Happyness[,-c(1,2)]),
direction="backward")

bcw.model$call
lm(formula = Score ~ `GDP per capita` + `Social support` +
    `Healthy life expectancy` + `Freedom to make life choices` +
    `Perceptions of corruption`, data = Happyness[, -c(1,
    2)])

# El orden para desechar:

# - Generosidad es la primera y unica variable que desecho
```

Mejores modelos

Es posible obtener los mejores modelos dependiendo de la cantidad de variables que queramos utilizar.

```
#install.packages("leaps")
library(leaps)

model_subset <- regsubsets(biggest,
                           data=Happyness[,-c(1,2)],method="exhaustive",nbest=1)

summary(model_subset)$which
  (Intercept) `GDP per capita` `Social support` `Healthy life expectancy`
1        TRUE             TRUE            FALSE                     FALSE
2        TRUE             TRUE            FALSE                     FALSE
3        TRUE             TRUE             TRUE                     FALSE
4        TRUE             TRUE             TRUE                      TRUE
5        TRUE             TRUE             TRUE                      TRUE
6        TRUE             TRUE             TRUE                      TRUE
  `Freedom to make life choices` Generosity `Perceptions of corruption`
1                          FALSE      FALSE                       FALSE
2                           TRUE      FALSE                       FALSE
3                           TRUE      FALSE                       FALSE
4                           TRUE      FALSE                       FALSE
5                           TRUE      FALSE                        TRUE
6                           TRUE       TRUE                        TRUE
```

Regresión Robusta

Una alternativa para controlar casos atípicos es ajustar una modelo lineal robusto. Los modelos lineales robustos utilizan criterios diferentes al de los mínimos cuadrados y ponderan la influencia de los casos atípicos, por lo que producen coeficientes y -sobre todo- errores estandar más confiables. Las regresiones robustas otorgan un peso a cada observación, generalmente ponderan con valores menores a uno a los casos atípicos, reduciendo su influencia.

Un buen ejercicio es comparar el modelo de regresión lineal normal con el robusto y ver cómo cambian las estimaciones de β para cada modelo. Si se observan grandes diferencias significa que los valores atípicos, outliers o palanca, están siendo muy influyentes.

Regresión Robusta

```
#regresion lineal normal OLS
RL<-lm(Score ~gdp+Socialsupp+Healthylife+Freedom+corruption,data=df)

library(MASS)

#regresion lineal robusta
RLR<-rlm(Score ~gdp+Socialsupp+Healthylife+Freedom+corruption,data=df)
```

Comparativa OLS - Robust linear

```
coef(RL)
(Intercept)         gdp   Socialsupp Healthylife     Freedom   corruption
  1.8688725   0.7454527    1.1180315   1.0840162   1.5340094    1.1175531

coef(RLR)
(Intercept)         gdp   Socialsupp Healthylife     Freedom   corruption
  1.8768471   0.7270432    1.0787122   1.0867740   1.5999074    1.5747303
```

Pasos generales en una RL

- Tener claro el objetivo del modelo: ¿Ajustar bien? ¿Parsimonia?
- Observar matriz de correlación lineal de las variables predictoras con la variable dependiente, además entre sí mismas
- Utilizar algún criterio para seleccionar variables, `forward aic`, `backward aic` o `regsubsets` en base a algún criterio (bic, r^2 ajustado, suma cuadrática residual, etcétera)
- Realizar gráficos de dispersión de las variables predictoras seleccionadas y la variable respuesta. Detectar patrones extraños (heterocedasticidad, asociación en el tiempo, etcétera)

Pasos generales en una RL

- Realizar modelo y evaluarlo: ¿tiene sentido el modelo obtenido?. Analizar también su comportamiento residual (supuestos de normalidad, homocedasticidad, independencia, etcétera), ¿qué tan bueno es el ajuste?
- Realizar una regresión robusta y ver si los parámetros estimados cambian mucho al compararlos con la regresión normal, en ese caso, realizar análisis de datos atípicos (quitarlos, evaluar transformación de variables, etcétera)
- Si a pesar de realizar tratamientos hay problemas severos en los supuestos o desempeño del modelo, evaluar otro tipo de modelos: modelos lineales generalizados, modelos no lineales, suavizamientos, etcétera

Verdadero o Falso

1. Las observaciones outliers o atípicas siempre hay que quitarlas al realizar un modelo de regresión lineal.
2. Siempre tiene sentido considerar un modelo de regresión lineal con el intercepto estimado.
3. El cálculo de $\hat{\beta}$ depende de las escalas de medición de las variables predictoras y se puede ver influenciado por observaciones atípicas.
4. La interpretación de $\hat{\beta}_1$ es exactamente igual en un modelo de regresión lineal simple o múltiple.
5. Una regresión lineal múltiple es una versión generalizada de la regresión lineal simple a p dimensiones, con p la cantidad de covariables.
6. El comportamiento de los residuos refleja el desempeño del ajuste del modelo de regresión lineal, además, de evidenciar qué tanto se cumplen los supuestos.

Respuestas

1. Falso. Es necesario cuantificar qué tanto afecta la observación outlier en la estimación de $\hat{\beta}$.
2. Falso. Pudiera pasar que el intercepto resulta negativo, pero que en el contexto, un intercepto negativo carezca de sentido.
3. Verdadero. El cálculo de $\hat{\beta}$ se obtiene minimizando la distancia de la recta a las observaciones, por lo tanto, se puede ver influenciado por las observaciones atípicas y por las escalas de medición de las variables.
4. Falso. En un modelo de regresión múltiple es necesario dejar fijas las demás variables para poder concluir en términos de la variable asociada a $\hat{\beta}_1$.
5. Verdadero. Corresponde a una recta generalizada a p dimensiones.
6. Verdadero. Es necesario evaluar el comportamiento de los residuos, para evaluar supuestos y calidad del ajuste.