Supuestos paramétricos

Predicción y residuales

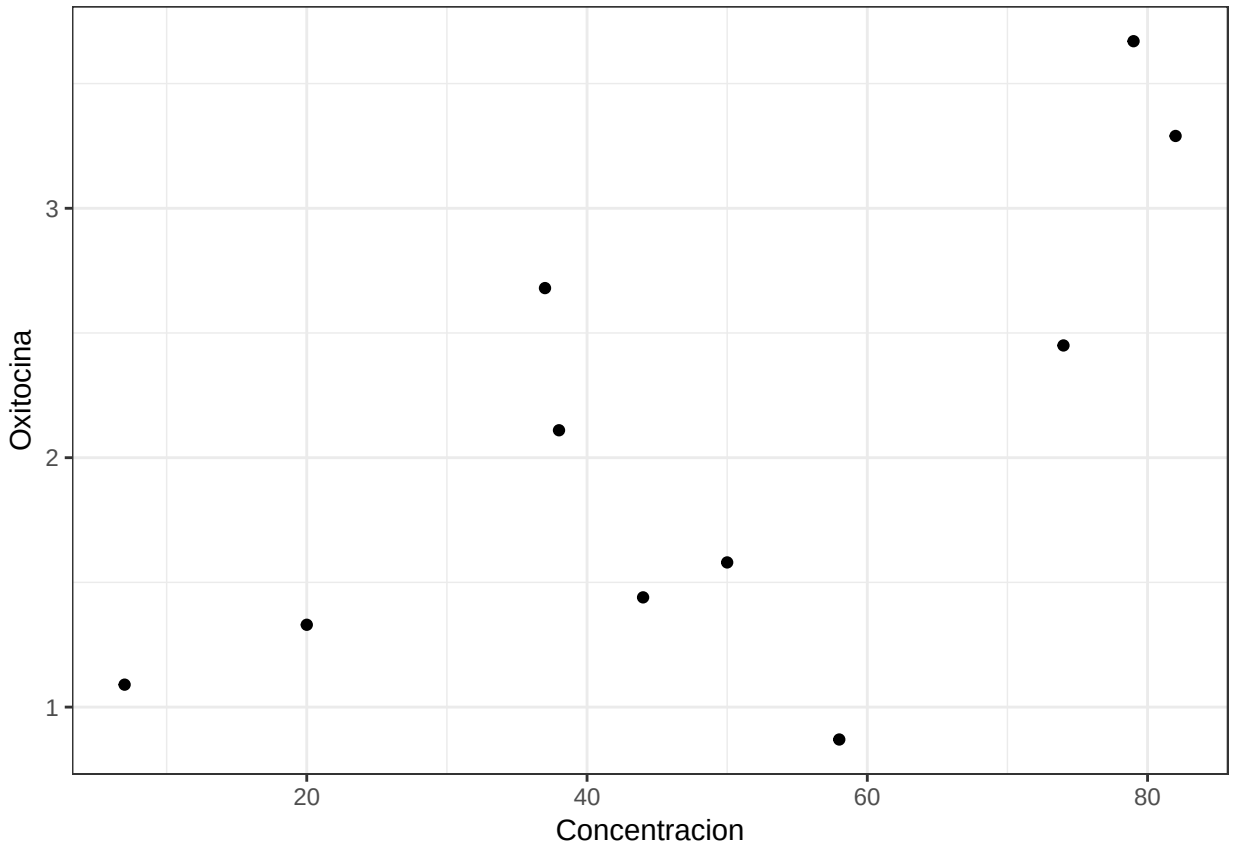
```
library(tidyverse)
ox_ob <- read.csv("oxitocina-observacion.csv" )
ox_ob <- mutate(ox_ob, Oxitocina = Oxytocin, Concentracion = Concentration) %>%
         select(ID, Concentracion, Oxitocina)
mod_ox_ob <- lm(Oxitocina ~ Concentracion, data = ox_ob)
ox_ob$prediccion <- predict(mod_ox_ob)
ox_ob$residuales <- residuals(mod_ox_ob)

head(ox_ob)
```

```
##   ID Concentracion Oxitocina prediccion  residuales
## 1  1            74      2.45   2.699549 -0.24954873
## 2  2            79      3.67   2.828742  0.84125829
## 3  3            58      0.87   2.286131 -1.41613121
## 4  4            38      2.11   1.769359  0.34064068
## 5  5            82      3.29   2.906257  0.38374251
## 6  6            20      1.33   1.304265  0.02573539
```

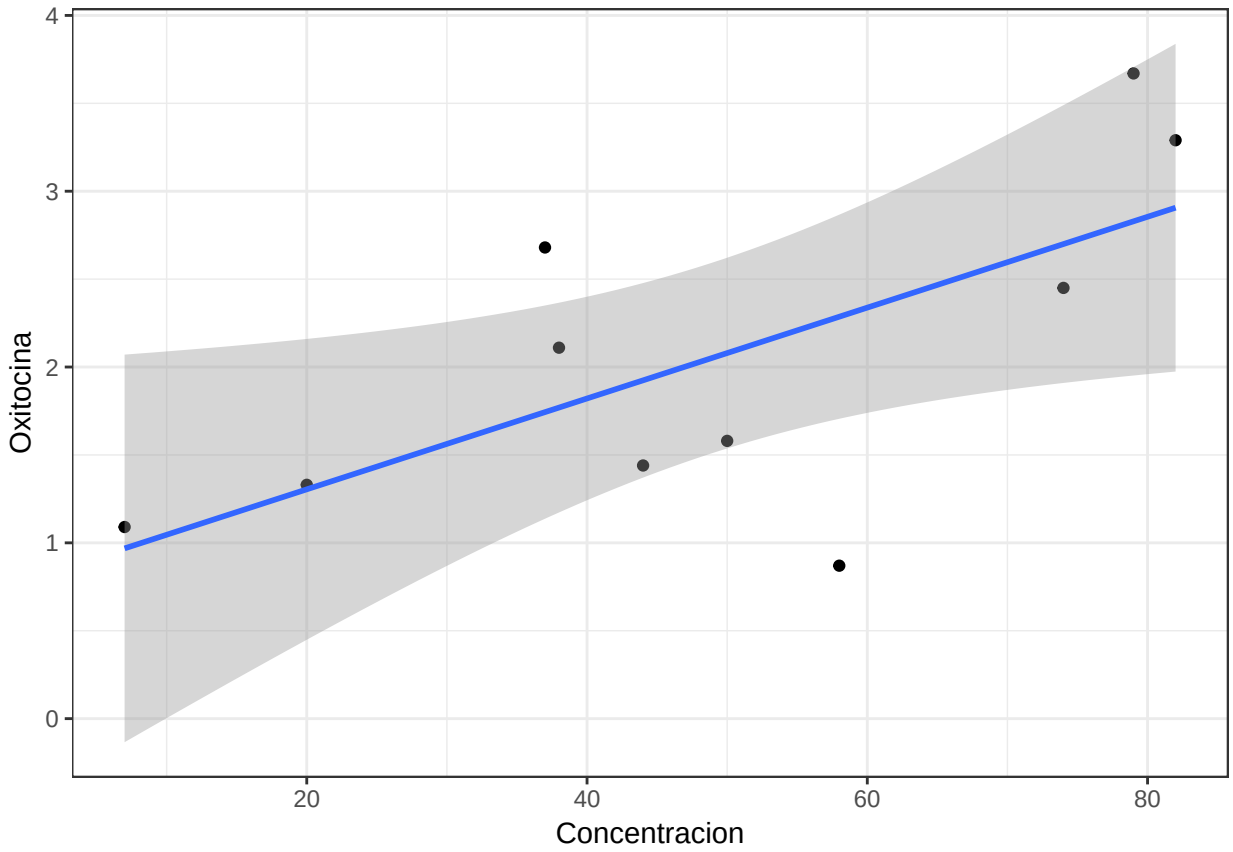
```
knitr::kable(ox_ob)
```

ID	Concentracion	Oxitocina	prediccion	residuales
1	74	2.45	2.6995487	-0.2495487
2	79	3.67	2.8287417	0.8412583
3	58	0.87	2.2861312	-1.4161312
4	38	2.11	1.7693593	0.3406407
5	82	3.29	2.9062575	0.3837425
6	20	1.33	1.3042646	0.0257354
7	7	1.09	0.9683629	0.1216371
8	37	2.68	1.7435207	0.9364793
9	44	1.44	1.9243909	-0.4843909
10	50	1.58	2.0794225	-0.4994225

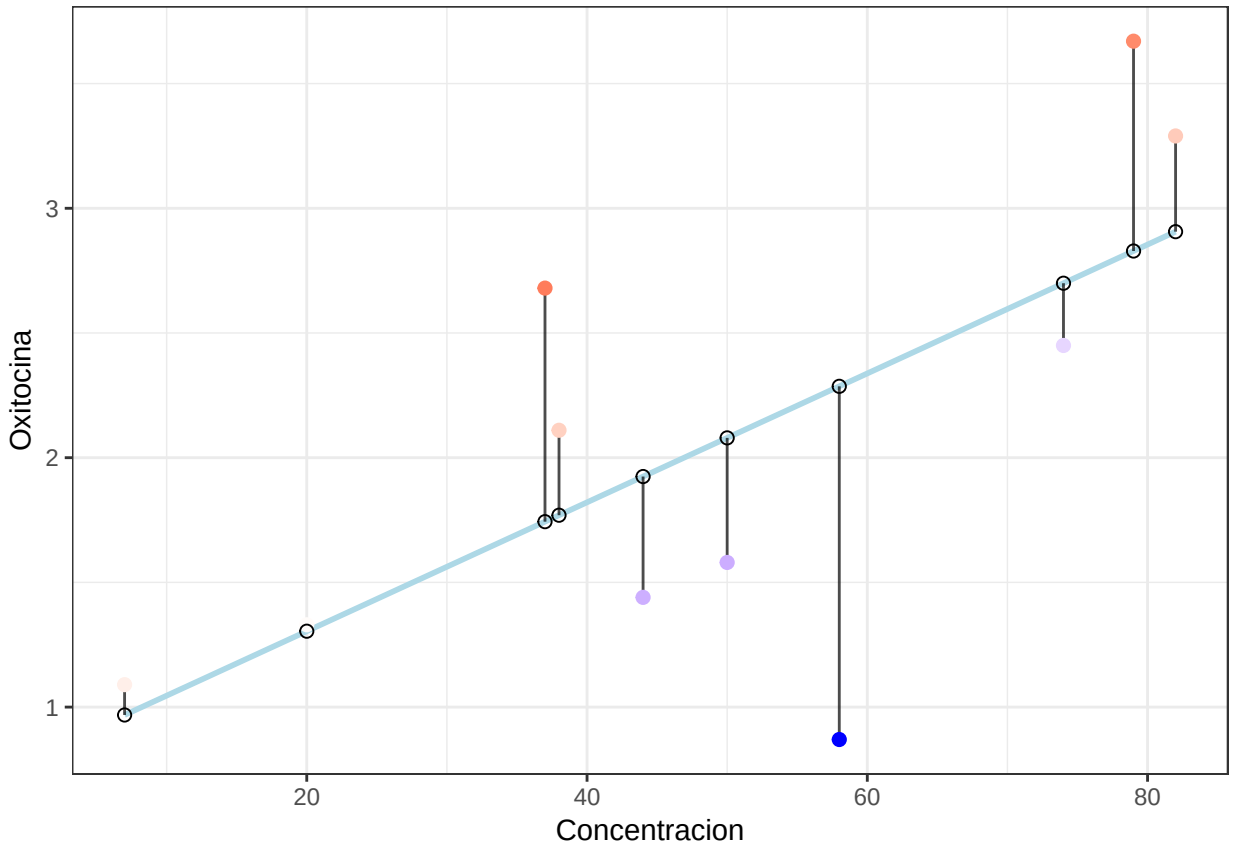
```
ggplot(ox_ob, aes(x = Concentracion, y = Oxitocina)) +
  geom_point() +
  theme_bw()
```




```
ggplot(ox_ob, aes(x = Concentracion, y = Oxitocina)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  theme_bw()
```




```
ggplot(ox_ob, aes(x = Concentracion, y = Oxitocina)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightblue") +
  geom_segment(aes(xend = Concentracion, yend = prediccion), alpha = .7) +
  geom_point(aes(color = residuales), size = 2) +
  scale_color_gradient2(low = "blue", mid = "white", high = "red") +
  guides(color = FALSE) +
  geom_point(aes(y = prediccion), shape = 1, size = 2) +
  theme_bw()
```




```
ggplot(ox_ob, aes(x = Concentracion, y = Oxitocina)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  theme_bw()
```

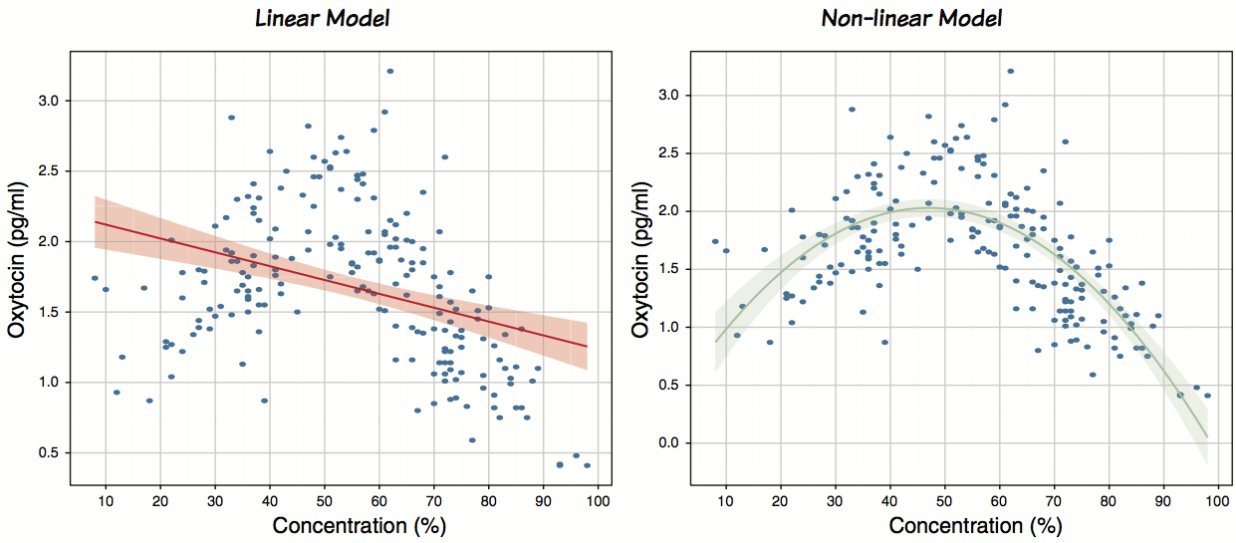
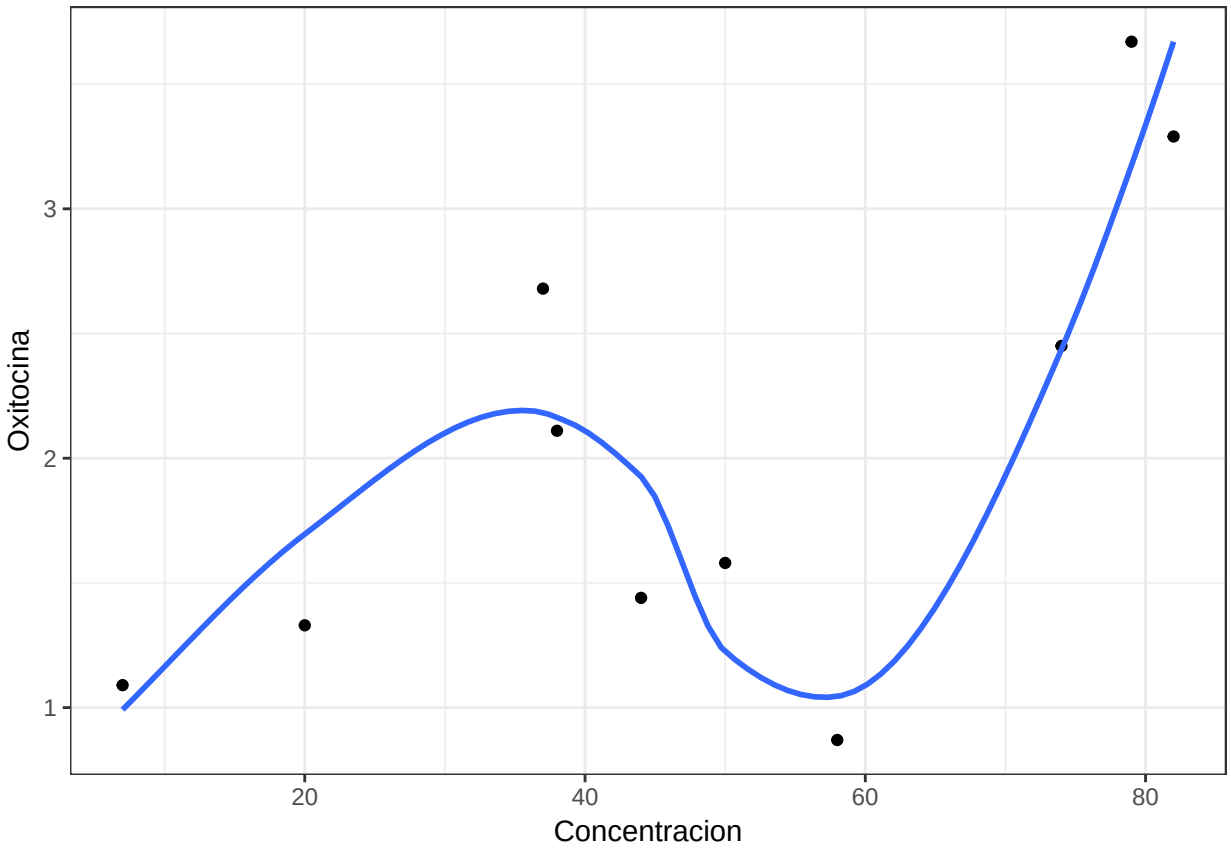


Figure 1:



Aditividad y linealidad

El supuesto de aditividad y linealidad es el más importante, porque es el supuesto de que el proceso del mundo real que quieres modelar puede ser capturado por la ecuación lineal. **Linealidad** se refiere a asumir

que la variable respuesta está linealmente relacionada a un predictor en particular (i.e., su relación puede ser capturada por una *línea recta*); mientras que **aditividad** se refiere al supuesto de que si tienes varios preddictores, entonces su efecto combinado es bien descrito *sumando* sus efectos. En otras palabras, significa que el proceso de interés es descrito con precisión por la siguiente ecuación.

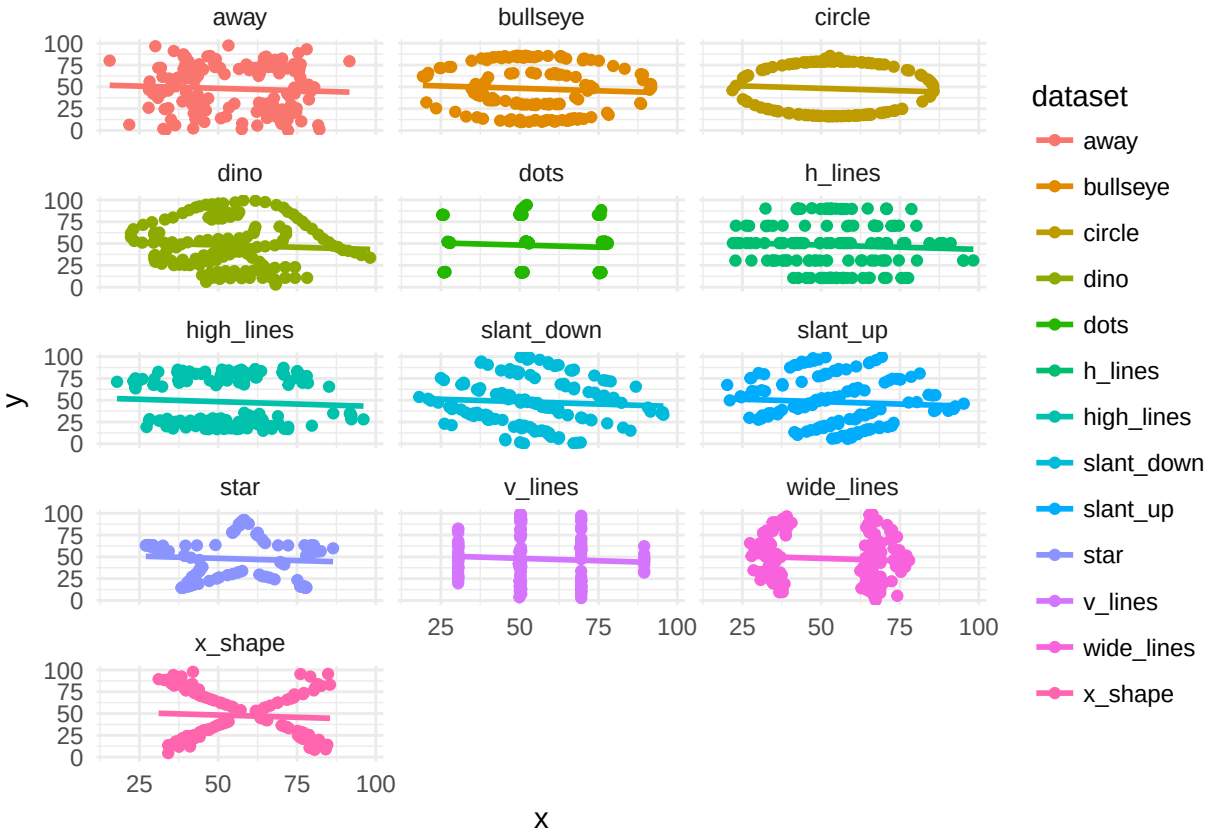
$$variablerespuesta_i = (b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_n X_{ni})$$

Si el proceso no sigue un patrón lineal, estás ajustando el modelo incorrecto.

¿Cómo lo pruebo?

Gráficamente

```
library(ggplot2)
library(datasauRus)
ggplot(datasaurus_dozen, aes(x = x, y = y, colour = dataset))+
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  theme(legend.position = "none")+
  facet_wrap(~dataset, ncol=3) +
  theme_minimal()
```



Independencia de los errores

Este supuesto implica que el $error_i$ para cada sujeto en la ecuación anterior no están relacionado uno con otro. Imagina que tu 'crush' (lo denominaremos **C**) y tú participan en una investigación donde tienen que

calificar cuánto les gusta una lista de canciones. Si les preguntamos estando juntos es muy probable que las calificaciones que den no sean independientes: Tú respuesta a cada cacnción dependerá de lo que responda 'C'. Sabemos que si intentamos predecir tus respuestas va a haber un error en las predicciones que hagamos. Crucialmente, debido a que tus respuestas y las de 'C' no son independientes, los errores (residuales) asociados tampoco lo serán; están violando el supuesto.

Ahoa imagina un escenario diferente: el investigador es más precavido y antes de pedirles calificar las canciones les coloca audífonos a prueba de sonido y los pone en dos habitaciones separadas. En este caso tus respuestas no se verían afectadas por las respuestas de 'C', por lo que su error debería de ser independiente: el error en predecir tus respuestas no debería ser influenciad por las respuestas de 'C'.

Genial... ¿Pero eso qué?

Es importante porque el cálculo del *error estándar* (la manera en que calculamos el error en un modelo) sólo es válida si las muestras son independientes y esta medida es utilizada para calacular los estadísticos y los intervalos de confianza: por lo que puede causar errores de inferencia en "cascada".

¿Cómo lo pruebo?

```
library(car)
modelo <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
durbinWatsonTest(modelo)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.100862      1.735915    0.31
##  Alternative hypothesis: rho != 0
```

Poner especial atención en la autocorrelación y el valor p.

Mis datos están autocorrelacionados ¿Qué hago?

La vida será más complicada y tendras que hacer uso de un tipo más avanzado de modelos conocidos como modelos multinivel.

Homosedasticidad

En un diseño donde mides varios grupos de participantes, este supuesto significa que cada una de las muestras proviene de pblaciones con la misma varianza.

¿Cómo lo pruebo?

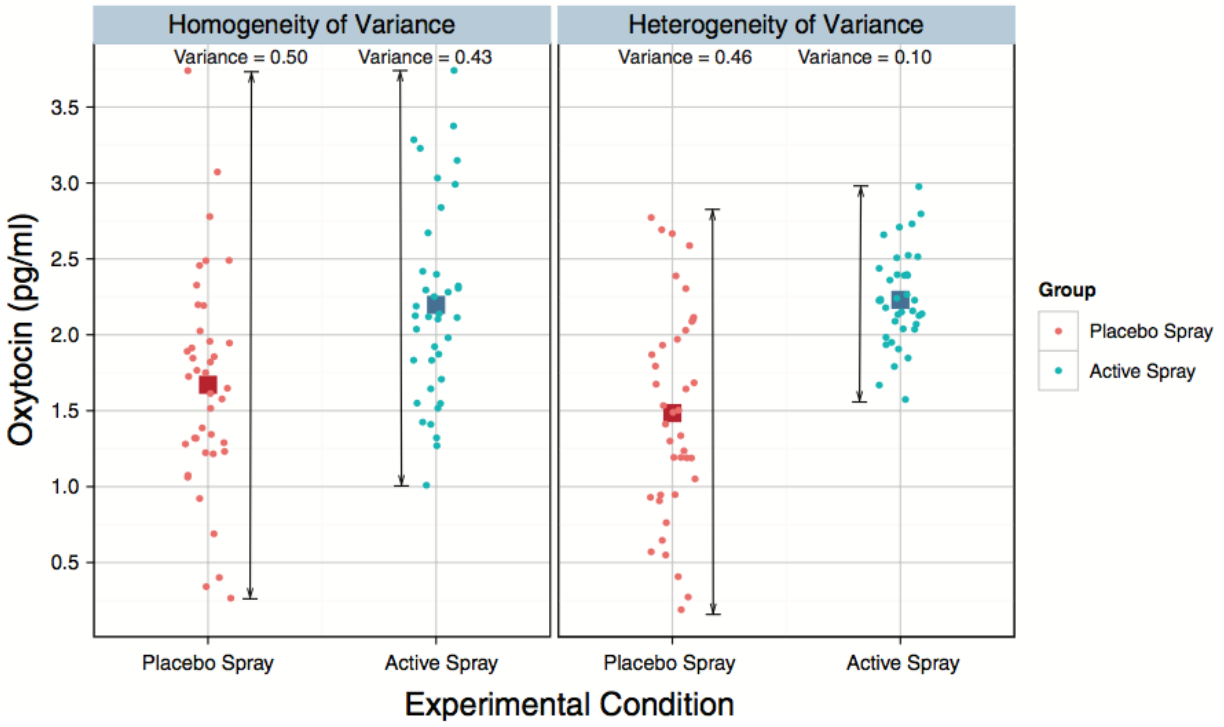


Figure 2:

Gráficamente

Estadísticamente

```
library(gvlma)
modelo <- lm(conformity ~ fcategory * partner.status, data = Moore)
leveneTest(modelo, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  5  1.7915 0.1373
##       39
```

```
gvmodel <- gvlma(modelo)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = conformity ~ fcategory * partner.status, data = Moore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6250 -2.9000 -0.2727  2.7273 11.3750
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       11.8571     1.7307   6.851 3.44e-08 ***
```

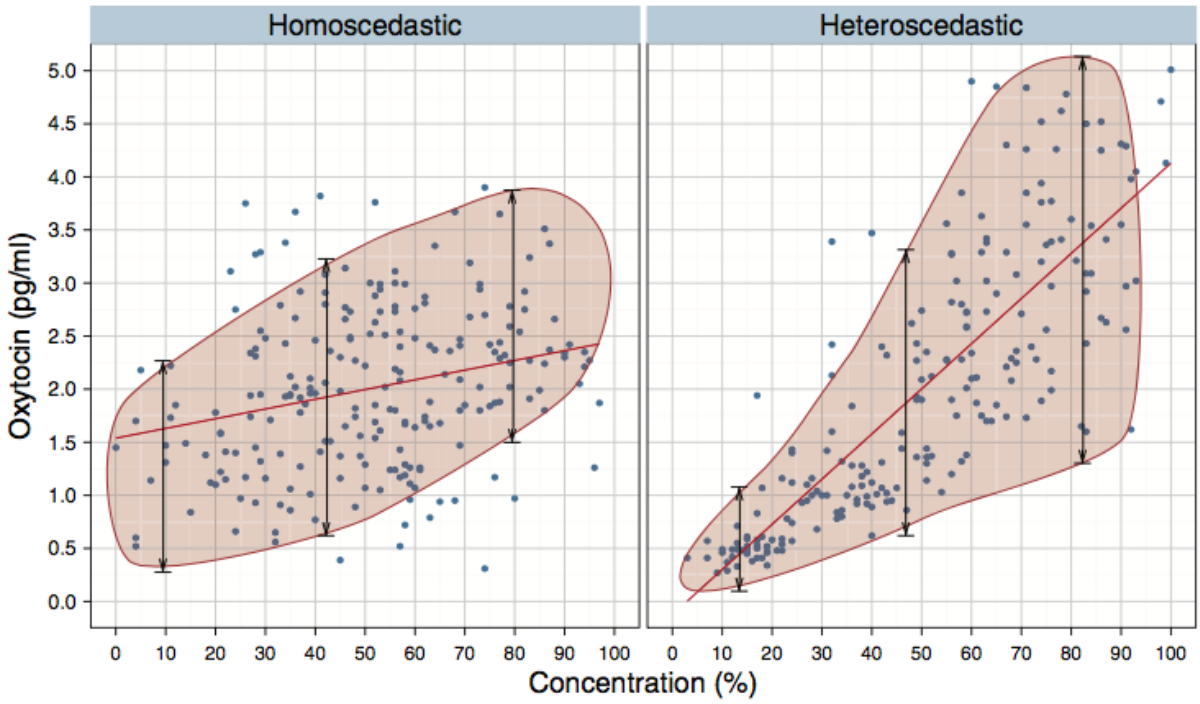


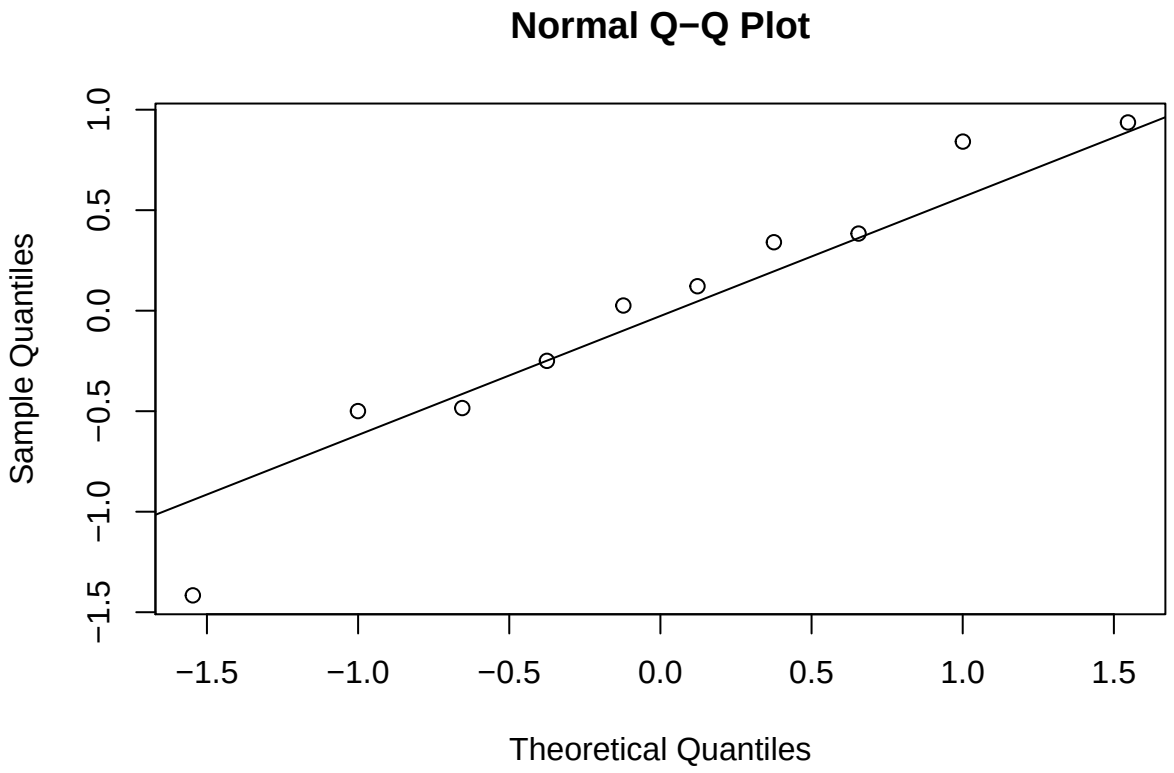
Figure 3:

```
## fcategorylow                          5.5429     2.6813   2.067   0.0454 *
## fcategorymedium                       2.4156     2.2140   1.091   0.2819
## partner.statuslow                     0.7679     2.3699   0.324   0.7477
## fcategorylow:partner.statuslow       -9.2679     3.4507  -2.686   0.0106 *
## fcategorymedium:partner.statuslow    -7.7906     3.5728  -2.181   0.0353 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.579 on 39 degrees of freedom
## Multiple R-squared:  0.3237, Adjusted R-squared:  0.237
## F-statistic: 3.734 on 5 and 39 DF,  p-value: 0.007397
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = modelo)
##
##                         Value p-value                Decision
## Global Stat        2.480e+00  0.6482 Assumptions acceptable.
## Skewness           1.624e+00  0.2025 Assumptions acceptable.
## Kurtosis           8.772e-02  0.7671 Assumptions acceptable.
## Link Function      1.419e-19  1.0000 Assumptions acceptable.
## Heteroscedasticity 7.683e-01  0.3807 Assumptions acceptable.
```

"Distribución Normal"

Gráficamente

```
mod_ox_ob <- lm(Oxitocina ~ Concentracion, data = ox_ob)
qqnorm(residuals(mod_ox_ob)); qqline(residuals(mod_ox_ob))
```



Estadísticamente

```
mod_ox_ob <- lm(Oxitocina ~ Concentracion, data = ox_ob)
shapiro.test(ox_ob$Oxitocina)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ox_ob$Oxitocina
## W = 0.93815, p-value = 0.5326
```

```
shapiro.test(residuals(mod_ox_ob))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_ox_ob)
## W = 0.95477, p-value = 0.7249
```

Multicolinealidad no perfecta

```
vif(modelo)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## fcategory                4.480519  2        1.454896
## partner.status           3.011905  1        1.735484
## fcategory:partner.status 7.548701  2        1.657555
```

Créditos: Gran parte de los ejemplos e imágenes son parte del libro: Field, A.(2016) *An Adventure in Statistics: The Reality Enigma*. UK: SAGE.