Ejercicio 2.1: Ejemplo outliers con Regresión

Silvia Pineda

Lectura de datos

```
cars_original <- cars # original data
### Introducimos unos outliers de forma manual
outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218))
cars_outliers <- rbind(cars_original, outliers)  # data with outliers.
```

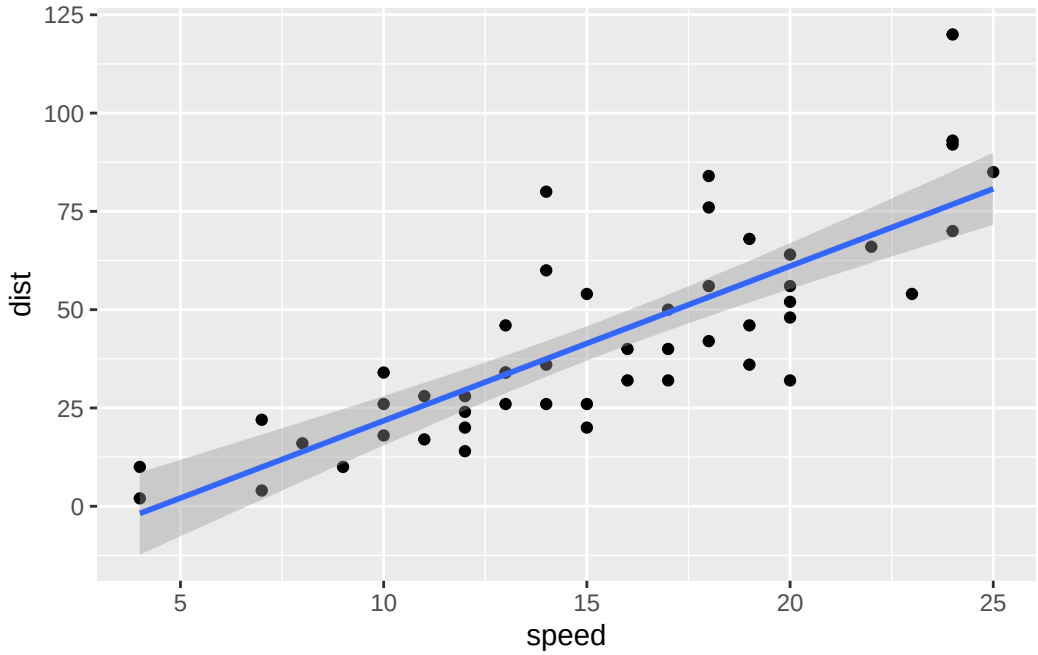
Ajustar recta de regresión con datos originales

```
summary(lm(cars_original$dist~cars_original$speed))
```

```
Call:
lm(formula = cars_original$dist ~ cars_original$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -17.5791     6.7584  -2.601   0.0123 *
cars_original$speed   3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
```

```
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
library(ggplot2)
ggp_original <- ggplot(cars_original,aes(speed, dist)) +
geom_point()


plot1<-ggp_original +
stat_smooth(method = "lm",
formula = y ~ x,
geom = "smooth")
plot1
```



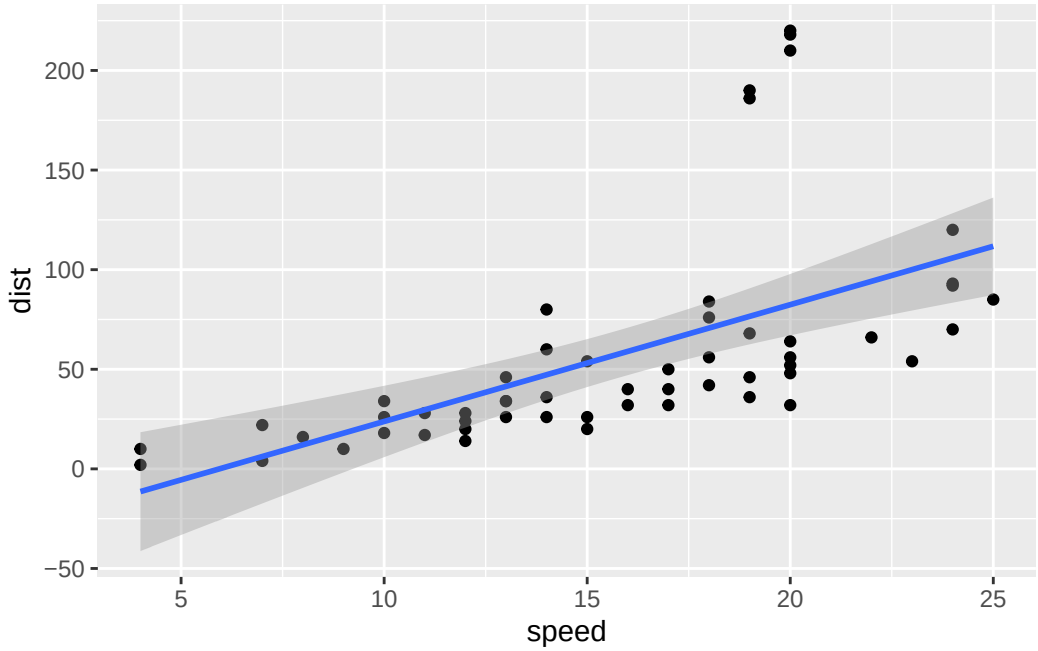
En esta regresión lineal, vemos un buen ajuste con un $R^2 = 0.65$ y un p-valor de la variable speed $= 1.49 * 10^{-12}$. Además en la recta de regresión vemos también un buen ajuste a cada uno de los puntos.

Ajustar recta de regresión con datos con outliers

```r
summary(lm(cars_outliers$dist~cars_outliers$speed))
```

```
Call:
lm(formula = cars_outliers$dist ~ cars_outliers$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-50.425 -26.583 -11.516   4.962 137.575

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -34.847     19.174  -1.817   0.0748 .
cars_outliers$speed    5.864      1.155   5.075 5.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.01 on 53 degrees of freedom
Multiple R-squared:  0.3271,    Adjusted R-squared:  0.3144
F-statistic: 25.76 on 1 and 53 DF,  p-value: 5.089e-06
```

```r
ggp_outlier <- ggplot(cars_outliers,aes(speed, dist)) +
  geom_point()

plot2<-ggp_outlier + stat_smooth(method = "lm",
                                 formula = y ~ x,geom = "smooth")

plot2
```

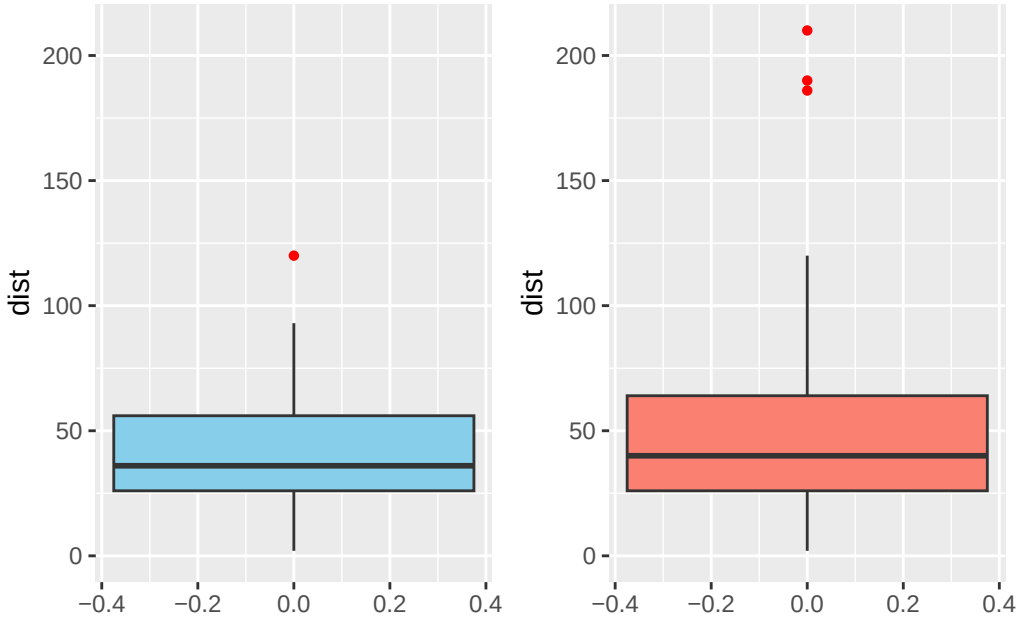


En este caso, vemos claramente los outliers fuera de la recta de regresión y vemos como tanto el R^2 como el nivel de significación han disminuido.

Diagramas de cajas para la detección de outliers

```
p1 <- ggplot(cars_original, aes(y = dist)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16) +
  scale_y_continuous(limits = c(0, 210))

p2 <- ggplot(cars_outliers, aes(y = dist)) +
  geom_boxplot(fill = "salmon", outlier.color = "red", outlier.shape = 16) +
  scale_y_continuous(limits = c(0, 210))

library(patchwork)
combined_plot <- p1 + p2 + plot_layout(ncol = 2)

# Mostrar el gráfico combinado
print(combined_plot)
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Aquí vemos como en la segunda bases de datos, los outliers se ven claramente muy alejados de los bigotes.