

# Primera tarea práctica: Regresión lineal.

Marcelo Molinatti

2023-03-20

## Regresión lineal simple.

Considere los datos experimentales de la tabla, que se obtuvieron de  $n = 33$  muestras de desechos tratados químicamente en un estudio realizado en *Virginia Tech*. Se registraron los valores de  $x$ , la reducción porcentual de los sólidos totales, y de  $y$ , el porcentaje de disminución de la demanda de oxígeno químico.

Al inspeccionar dicho diagrama se observa que los puntos se acercan mucho a una línea recta, lo cual indica que la suposición de linealidad entre las dos variables parece ser razonable y el `lm` modelo seleccionado para el análisis.

```
ggplot(df_RLS, aes(x=solid_reduction, y=oxigen_reduction)) +  
  geom_point(size=1.8) +  
  geom_smooth(method=lm, se=FALSE) +  
  xlab("Fuerza del Brazo") + ylab("Levantamiento dinámico") +  
  theme_light()
```

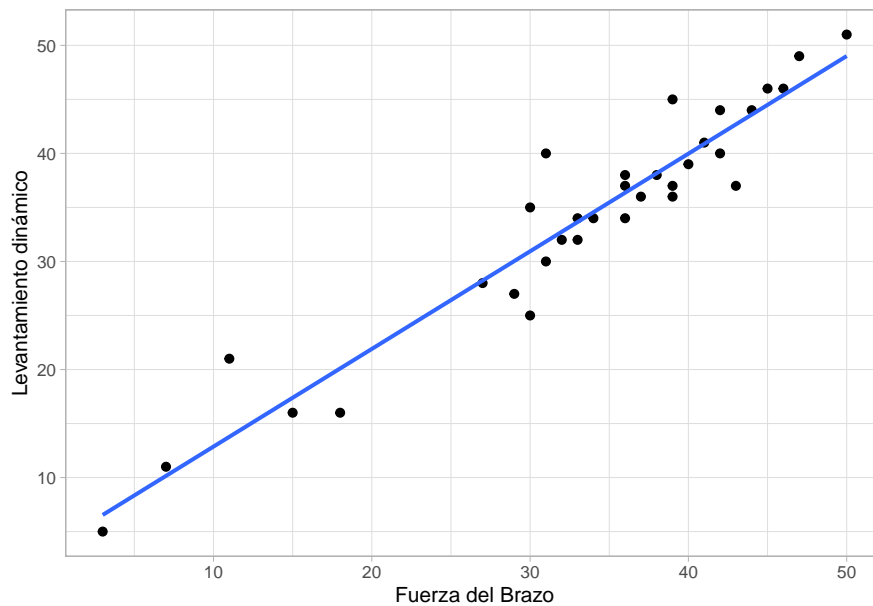


Figura 1: Reduccion de la demanda de oxigeno con respecto a la reduccion de solidos totales.

Ajustamos un modelo de regresión lineal simple de la forma  $y = \beta_0 + \beta_1 x$ , para poder estimar los coeficientes del modelo de regresión  $\beta_0$  y  $\beta_1$

```
fit <- lm(oxigen_reduction ~ solid_reduction, df_RLS)
```

```
as_tibble(summary(fit)$coefficients) |>
  tibble::add_column(term=c("Reduccion de Solidos", "Residuales"), .before=1) |>
  kable(digits=3, row.names=FALSE, booktabs=TRUE,
    col.names=c("", "Estimado,  $\hat{\beta}$ ", "Desv. Estandar", " $\hat{t}$ ", " $P(t > \hat{t})$ "),
    caption="Resultados de la regresion lineal simple.",
    escape=FALSE, label="pRLS-01", format.args = list(big.mark=".")) %>%
  kable_classic(position = "center", latex_options = "hold_position")
```

Tabla 1: Resultados de la regresion lineal simple.

	Estimado, $\hat{\beta}$	Desv. Estandar	$\hat{t}$	$P(t > \hat{t})$
Reduccion de Solidos	3,830	1,768	2,166	0,038
Residuales	0,904	0,050	18,030	0,000

Como se observa, ambos estimadores (del coeficiente y de la pendiente) son significativamente distintos de cero, dado el valor de probabilidad asociado para el estadístico  $\hat{t}$  que indica que una desviación tan grande solo por azar es improbable. Este estadístico busca contrastar las hipótesis  $H_0 : \beta_i = 0$  contra la alternativa  $H_1 : \beta_i \neq 0$  para cada  $i = 0, 1$ . Como la probabilidad asociada es muy pequeña, se concluye que la probabilidad de que  $\hat{\beta}_i \sim N(0, \sigma_{\beta_i})$  es muy baja y que estos deben venir de alguna otra distribución con media distinta de cero.

Otra forma de validar el modelo de regresión es comparando el modelo planteado con un modelo nulo que no incluye pendiente (es decir, que no incluye la relación lineal con la variable independiente  $x$ ), utilizando una prueba  $F$  como se muestra en la tabla 2.

```
broom::tidy(aov(fit)) |>
  mutate(term=c("Reduccion de Solidos", "Residuales")) |>
  kable(digits=3, row.names=FALSE, booktabs=TRUE,
    col.names=c("", "df", "SS", "MS", " $\hat{F}$ ", " $P(F > \hat{F})$ "),
    caption="\\label{tab:pRLS-04}Tabla ANOVA para validar el modelo de regresion.",
    escape=FALSE, format.args = list(big.mark=".")) %>%
  kable_classic(position = "center", latex_options = "hold_position")
```

Tabla 2: Tabla ANOVA para validar el modelo de regresion.

	df	SS	MS	$\hat{F}$	$P(F > \hat{F})$
Reduccion de Solidos	1	3.390,551	3.390,551	325,08	0
Residuales	31	323,327	10,430		

De la tabla se observa que al comparar las varianza del modelo con pendiente y el modelo nulo (solo intercepto), se obtiene un valor de  $F$  mucho mas grande que el que se esperaría por azar, y es por ello que se le asocia una probabilidad aproximadamente nula (a 4 espacios decimales, el valor es cero). Esto no da la confianza de escoger el modelo de regresión como un buen modelo de trabajo que permite describir los datos observados, y realizar predicciones. De hecho, es posible calcular y obtener una medida de asociación entre las variables,  $R^2$ , cuyo valor es 0,9101324, el cual nos permite concluir que el modelo es capaz de explicar un 91 % de la varianza observada.

Sin embargo, aun es necesario validar las suposiciones del modelo de regresión, verificando lo valores residuales. Los gráficos mostrados en la figura 2 muestran que los residuales no tienen un comportamiento normal: es fácil ver del *QQ-plot* que varias observaciones se desvían mas de los esperado, tanto por encima como por debajo. El gráfico de residuales a la izquierda muestra que aun se percibe cierto grado de linealidad entre los valores esperados y los residuales, y permite constatar la presencia de al menos dos residuales que se desvían mas de dos desviaciones estándar de la media de cero. El tercer gráfico (a la derecha) muestra la información del gráfico de residuales, pero donde cada punto se ha escalado en tamaño usando como factor la

influencia que cada punto tiene sobre la estimación. Se puede observar, que las observaciones asociadas a valores grandes o pequeños de sólidos totales tienen la mayor influencia y que uno de los atípicos tiene una gran influencia en la recta estimada.

```
augmented_fit <- broom::augment(fit)

cowplot::plot_grid(
  ggplot(augmented_fit, aes(x=.fitted, y=.std.resid)) +
    geom_point(size=1.8) +
    xlab("Predichos") + ylab("Residuales") +
    geom_hline(yintercept=c(0, -1.96, 1.96), linetype=c(1, 2, 2)) +
    theme_light(),
  ggplot(augmented_fit, aes(sample = .resid)) +
    stat_qq() +
    stat_qq_line(color="blue") +
    theme_light(),
  ggplot(mutate(augmented_fit, .too.large=ifelse(abs(.std.resid) > 1.96, 1, 0)),
    aes(x = .fitted, y = .std.resid)) +
    geom_point(aes(colour = .too.large, size=.hat * 100), alpha = .8) +
    geom_hline(yintercept = c(1.96, 3), linetype = 2, col = c("gray60", "gray50")) +
    xlab("Predichos") + ylab("Residuales Estandarizados") +
    theme_bw() +
    theme(legend.position="none") +
    geom_text(aes(label=ifelse(.std.resid > 1.96, as.character(paste(solid_reduction, oxigen_reduct.
      hjust=0, vjust=0),
    nrow=1)
```

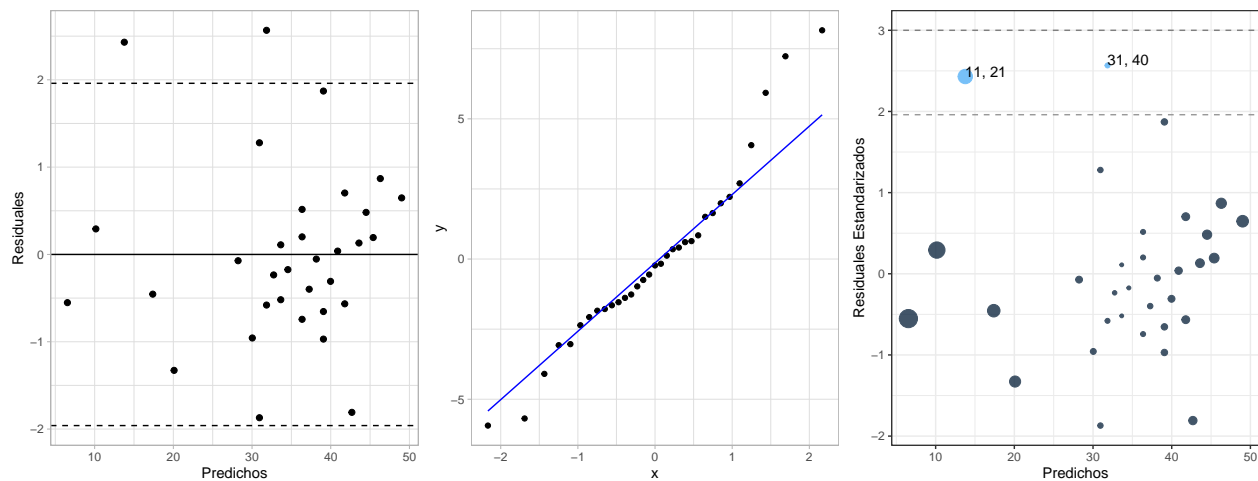


Figura 2: Graficos de residuales para validacion del modelo de regresion.

Dado estos resultados, se concluye que el problema no parece ser que el modelo sea incorrecto, sino que no se esta tomando en cuenta la influencia que cada observación (particularmente los atípicos) tienen sobre la estimación de los coeficientes. Es por ello que, en lugar de una transformación de los datos, se prefirió realizar una regresión de mínimos cuadrados ponderados, usando como pesos la varianza residual estimada de cada una de las observaciones:

$$w_i = \frac{1}{\hat{\sigma}^2(1 - h_{ii})}$$

donde  $h_{ii}$  es la palanca (*leverage*) de la  $i$ -ésima observación. Realizando la regresión ponderada usando estos pesos arroja los resultados mostrados en la tabla 3, donde se observa el cambio ligero en los coeficientes (los cuales aun siguen siendo significativos): aumento de la pendiente y caída del intercepto.

```
fit2 <- lm(oxigen_reduction ~ solid_reduction, df_RLS, weights=1 / (broom::glance(fit)$sigma ** 2 * (1 - h_ii)))
as_tibble(summary(fit2)$coefficients) |>
  tibble::add_column(term=c("Reduccion de Solidos", "Residuales"), .before=1) |>
  kable(digits=3, row.names=TRUE,
    col.names=c("", "Estimado",  $\hat{\beta}$ , "Desv. Estandar",  $\hat{\sigma}$ , " $P(t > \hat{t})$ "),
    escape=FALSE, label="tab:pRLS-03", format.args = list(big.mark="."),
    caption="\\label{tab:pRLS-03}Resultados de la regresión lineal ponderada.") %>%
  kable_classic(position = "center", latex_options = "hold_position")
```

Tabla 3: Resultados de la regresión lineal ponderada.

		Estimado, $\hat{\beta}$	Desv. Estandar	$\hat{\sigma}$	$P(t > \hat{t})$
1	Reduccion de Solidos	3,869	1,672	2,314	0,027
2	Residuales	0,903	0,048	18,900	0,000

El modelo nuevo, sigue siendo significativo al compararlo con el modelo nulo, e incluso el error cuadrado medio es 10 veces menor en el modelo de regresión ponderado (resultados no mostrados). La superposición de la nueva recta permite darnos cuenta que la mejora no es muy grande, y que la ponderación solo resulta en un aumento de 0,744 % en la varianza explicada.

```
ggplot(df_RLS, aes(x=solid_reduction, y=oxigen_reduction)) +
  geom_point(size=1.8) +
  geom_smooth(method=lm, se=FALSE) +
  geom_abline(slope=coef(fit2)[[2]], intercept=coef(fit2)[[1]]) +
  xlab("Fuerza del Brazo") + ylab("Levantamiento dinámico") +
  theme_light()
```

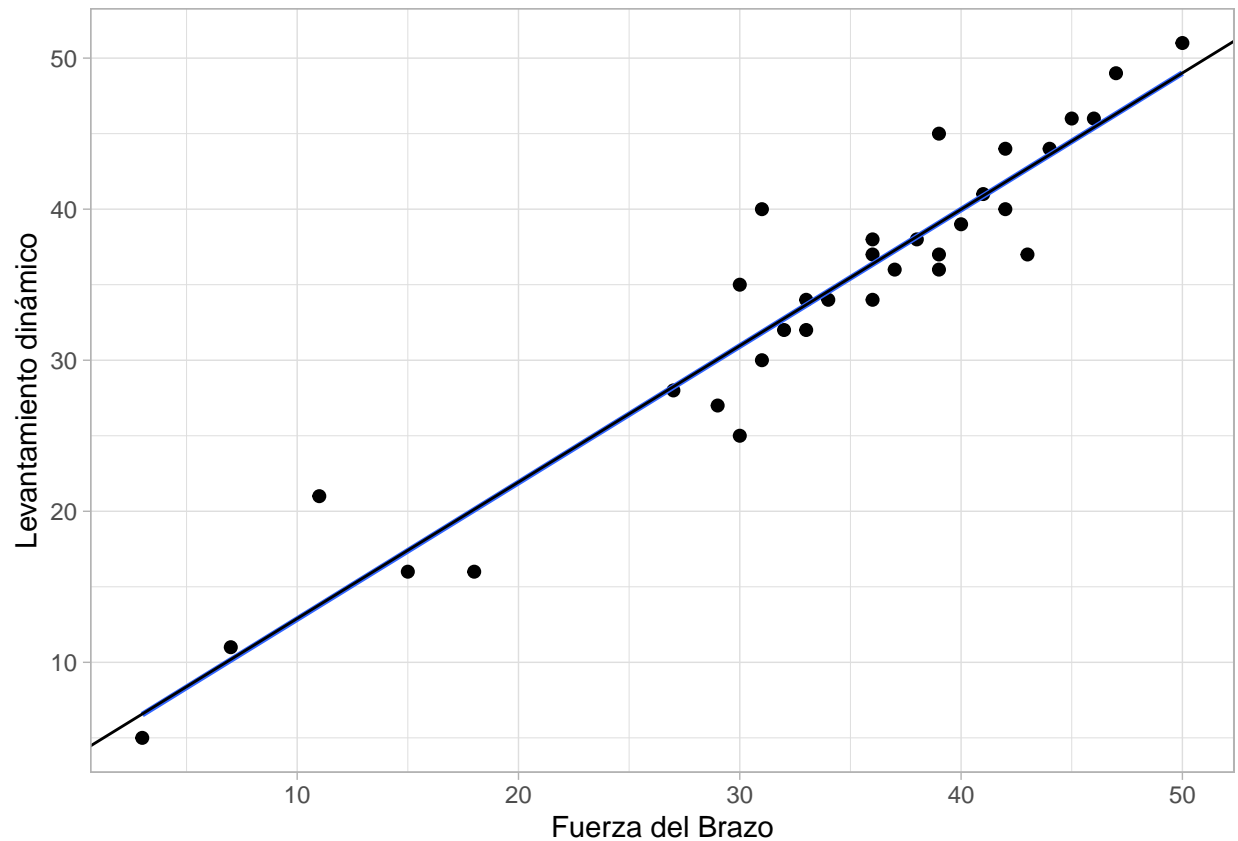


Figura 3: Reduccion de la demanda de oxigeno con respecto a la reduccion de solidos totales.

Dado que no hubo una mejora importante, seria adecuado considerar una regresion a trozos. Una inspeccion mas cuidadosa de los datos parece indicar que el conjunto de datos se podria dividir en dos regiones, cuyo paso de una region a otra depende de si la cantidad de solidos totales es mayor o menor a aproximadamente 22.

```
sessionInfo()
```

```
## R version 4.2.2 Patched (2022-11-10 r83330)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=es_VE.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=es_VE.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=es_VE.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_VE.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
```

```
##
## other attached packages:
## [1] dplyr_1.0.10      cowplot_1.1.1      broom_1.0.1        ggplot2_3.3.6
## [5] kableExtra_1.3.4
##
## loaded via a namespace (and not attached):
## [1] tinytex_0.42      tidyselect_1.2.0   xfun_0.34          purrr_0.3.5
## [5] splines_4.2.2     lattice_0.20-45    colorspace_2.0-3   vctrs_0.5.0
## [9] generics_0.1.3    htmltools_0.5.3    viridisLite_0.4.1  yaml_2.3.6
## [13] mgcv_1.8-41       utf8_1.2.2         rlang_1.0.6        pillar_1.8.1
## [17] glue_1.6.2        withr_2.5.0        DBI_1.1.3          lifecycle_1.0.3
## [21] stringr_1.4.1     munsell_0.5.0      gtable_0.3.1       rvest_1.0.3
## [25] evaluate_0.17     labeling_0.4.2     knitr_1.40         fastmap_1.1.0
## [29] fansi_1.0.3       highr_0.9          scales_1.2.1       backports_1.4.1
## [33] webshot_0.5.4     farver_2.1.1       systemfonts_1.0.4  digest_0.6.30
## [37] stringi_1.7.8     grid_4.2.2         cli_3.4.1          tools_4.2.2
## [41] magrittr_2.0.3    tibble_3.1.8       tidyr_1.2.1        pkgconfig_2.0.3
## [45] Matrix_1.5-1      xml2_1.3.3         assertthat_0.2.1   rmarkdown_2.17
## [49] svglite_2.1.0     httr_1.4.4         rstudioapi_0.14    R6_2.5.1
## [53] nlme_3.1-162      compiler_4.2.2
```