EST-25134: Aprendizaje Estadístico

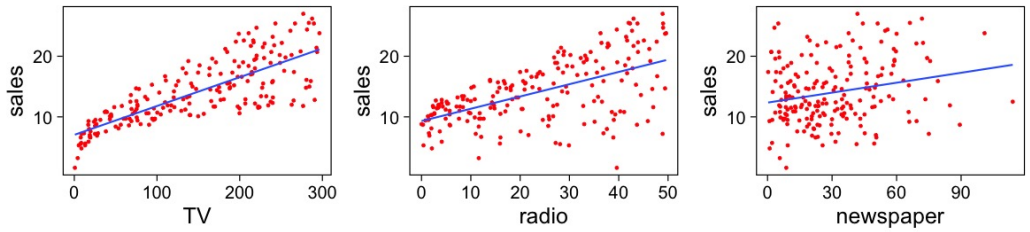
Profesor: Alfredo Garbuno Iñigo — Primavera, 2023 — Regresión lineal.
Objetivo. Repasaremos los conceptos de regresión lineal desde un punto de vista de inferencia. Haremos conexiones interesantes con conceptos clave en probabilidad y teoria de la información. Veremos métricas de desempeño para modelos de regresión. Nos llevará a cuestionar el modelo bajo el enfoque de predicción.
Lectura recomendada: Capítulo 3 de James et al. [1].

1. INTRODUCCIÓN

Aproximamos $f(X)$ como una combinación lineal de las entradas.

1.1. Datos de marketing

¿Qué preguntas serían las que nos interesarían?



¿Hay alguna relación entre datos de entrada? En especial, ¿hay alguna relación entre cuánto se le dedica al presupuesto de *marketing* y las ventas observadas? ¿Qué tipo de *marketing* contribuye mas a las ventas? ¿Qué tan precisos podemos ser con nuestras predicciones? La relación es lineal? ¿Hay algún tipo de sinergia?

2. EL MODELO SIMPLE

$$Y = \beta_0 + \beta_1 X + \varepsilon\,. \tag{1}$$

Si tuvieramos un estimador $\hat{\beta}$, ¿cómo podemos realizar predicciones?

2.1. Estimación de parámetros

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i\,. \tag{2}$$

Formulación de problema de optimización. Conexiones interesantes.

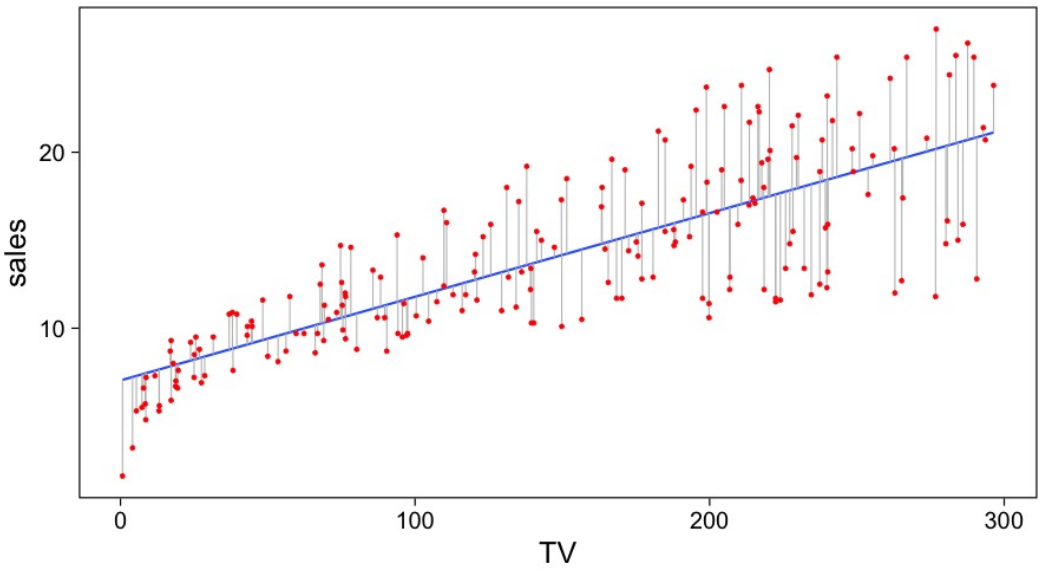


FIGURA 1. *Ajuste y residuales a la recta de mínimos cuadrados.*

2.2. Solución

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \tag{3}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,. \tag{4}$$

Qué deficiencia encuentras en el ajuste?

2.3. ¿Precisión en los estimadores?

$$\mathsf{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \tag{5}$$

$$\mathsf{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{6}$$

¿A qué se debe esta variabilidad? Se pueden construir **intervalos de confianza**.

```
model ▷
      summary()
```

LISTING 1. *Resumen del modelo.*

```

Call:
lm(formula = sales ~ TV, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.386 -1.955 -0.191   2.067   7.212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.03259    0.45784    15.4   <2e-16 ***
TV           0.04754    0.00269    17.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 3.26 on 198 degrees of freedom
Multiple R-squared:  0.612,     Adjusted R-squared:  0.61
F-statistic:  312 on 1 and 198 DF,  p-value: <2e-16
```

```
model ▷
  broom::tidy()
```

Listing 2. *Resumen del modelo (*tidy*).*

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   7.03     0.458        15.4 1.41e-35
2 TV            0.0475   0.00269      17.7 1.47e-42
```

```
genera_datos ← function(id){
  a ← 1; b ← 0; n ← 100
  tibble(x = runif(n, -1, 1),
         y = a * x + b + rnorm(n, sd = 1))
}
ajusta_modelo ← function(datos){
  modelo ← lm(y ~ x, datos)
  modelo
}
```

```
simulacion ←  tibble(id = seq(1, 10)) ▷
    mutate(datos  = map(id, genera_datos),
           modelo = map(datos, ajusta_modelo),
           ajuste = map(modelo, broom::tidy))
```

2.4. Prueba de hipótesis

$$H_0: \quad \text{No hay relación entre } X \text{ y } Y\,, \tag{7}$$

$$H_1: \quad \text{Existe una hay relación entre } X \text{ y } Y\,. \tag{8}$$

La prueba de hipótesis se efectúa en el contexto del modelo que estamos proponiendo.

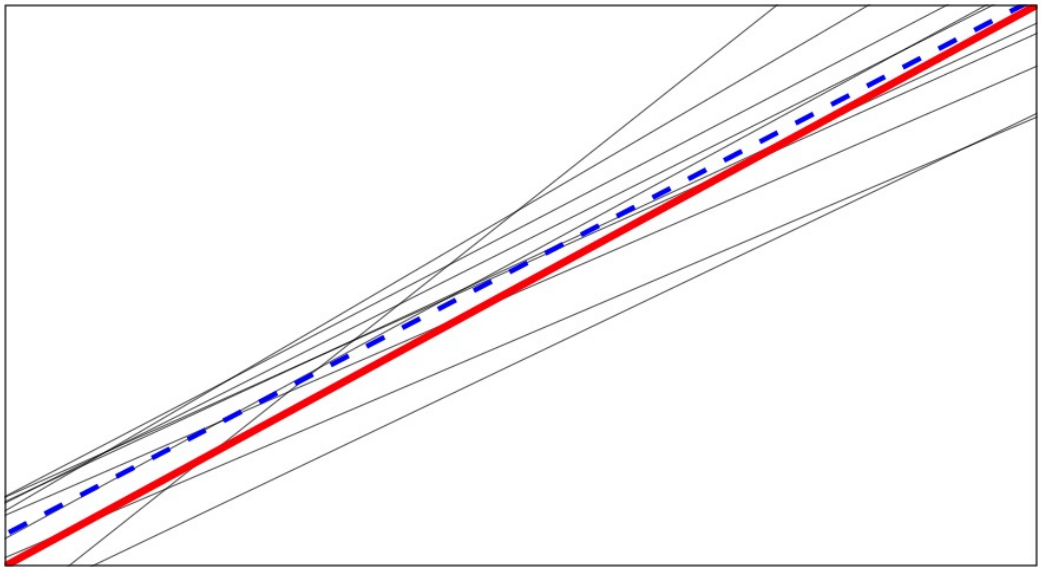


FIGURA 2. *Simulación de ajuste (variación en datos).*

2.5. El valor-p

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)}, \qquad \text{distribución } t_{n-2}\,. \tag{9}$$

2.6. Midiendo la precisión del modelo

$$\mathsf{RSE} = \sqrt{\frac{1}{n-2}\mathsf{RSS}}\,. \tag{10}$$

$$\mathsf{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,.$$

Es una métrica usual de ajuste. Nos dice qué tan precisos podemos ser al calcular una nueva predicción. Tiene sentido cuando comparamos con las unidades de la variable respuesta. Es decir, no es una métrica absoluta.

$$R^2 = \frac{\mathsf{TSS} - \mathsf{RSS}}{\mathsf{TSS}}\,. \tag{11}$$

$$\mathsf{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \,.$$

Hay que tener cuidado pues la R^2 es una métrica de correlación lineal, no de ajuste. Esto lo podemos ver en el caso sencillo de una variable respuesta y un modelo lineal. También está sujeta a la variación de la respuesta. Depende la aplicación para determinar cuándo tenemos un buen coeficiente de variación explicada.

3. EL MODELO MULTIVARIADO

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \,. \tag{12}$$

Cuando tenemos multiples predictores nos gustaría poder entender la relación de cada uno con la respuesta. ¿Ajustaríamos un modelo independiente con sólo un predictor?

3.1. Interpretación

$$\mathsf{sales} = \beta_0 + \beta_1 \times \mathsf{TV} + \beta_2 \times \mathsf{radio} + \beta_3 \times \mathsf{newspaper} + \varepsilon \,. \tag{13}$$

El modelo de regresión lineal usualmente se interpreta como los efectos (¿promedio, esperados?) de cada variable al mantener todas las demás *constantes*. Hay problemas cuando hay correlación entre predictores. Cuidado con datos observacionales.

3.2. Estimación

```
model ← lm(sales ~ ., data)
```

```
model ▷
  broom::tidy()
```

```
# A tibble: 4 × 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  2.94       0.312       9.42  1.27e-17
2 TV           0.0458     0.00139    32.8   1.51e-81
3 radio        0.189      0.00861    21.9   1.51e-54
4 newspaper   -0.00104    0.00587    -0.177 8.60e- 1
```

Desarollo de verosimilitud.

3.3. ¿Existe una relación entre la respuesta y los predictores?

Nos preguntamos si es que existe alguna $\beta_j \neq 0$.

$$F = \frac{(\mathsf{TSS} - \mathsf{RSS})/p}{\mathsf{RSS}/(n-p-1)} \sim F_{p,n-p-1}\,. \tag{14}$$

La prueba de hipótesis que formularíamos sería probar contra alguna $\beta_j \neq 0$. Se puede probar que si el supuesto del modelo lineal es correcto y bajo la hipótesis nula el cociente será cercano a 1. En caso de que la hipótesis `alternativa` sea cierta entonces $F > 1$.

```
model ▷
  broom::glance() ▷
  select(statistic, p.value, df, df.residual)
```

Listing 3. *Resumen global del modelo (*tidy*).*

```
# A tibble: 1 × 4
  statistic  p.value    df df.residual
      <dbl>    <dbl> <dbl>       <int>
1      570. 1.58e-96     3         196
```

- ¿Por qué tenemos que evaluar en conjunto?

¿Qué pasa en el caso con 100 predictores donde no hay relación?

3.4. ¿Cuáles son los predictores importantes?

Métodos de selección.

La idea mas ingenua es ajustar todas las posibles combinaciones. Pero se pueden construir modelos de manera secuencial . Usualmente ajustando y comparando con respecto a *alguna métrica*. Mas adelante lo estudiaremos.

3.5. ¿Qué tan bien ajusta el modelo?

Podemos usar las métricas típicas como el RSE o la R^2.

R^2: Agregar predictores siempre ayuda (en datos de entrenamiento).
RSE: Podemos tener problemas pues mientras mas variables agregemos si el cambio en residuales es pequeño en relación al aumento de p.

3.6. ¿Cómo predecimos y que tan precisa es nuestra predicción?

Podemos utilizar `intervalos confianza`. Mejor aún, podemos utilizar `intervalos de predicción`.

4. EXTENSIONES

4.1. Predictores cualitativos

Modelo con respuestas binarias (1D). ¿Qué tal que tenemos mas categorias?

4.2. Interacciones

Eliminar el `supuesto aditivo`: *interacciones* y *no-linealidad*.

```
model.1 ← lm(sales ~ TV + radio, data)
model.2 ← lm(sales ~ TV + radio + TV:radio, data)
```

Listing 4. *Ajuste de modelos sin/con interacciones.*

```
tibble(modelo = list(model.1, model.2),
       tipo   = c("lineal", "interaccion")) ▷
  mutate(resultados = map(modelo, broom::tidy)) ▷
  select(-modelo) ▷
  unnest(resultados) ▷
  select(tipo, term, estimate, p.value)
```

```
# A tibble: 7 × 4
  tipo        term        estimate  p.value
  <chr>       <chr>          <dbl>    <dbl>
1 lineal      (Intercept)  2.92    4.57e-19
2 lineal      TV           0.0458  5.44e-82
3 lineal      radio        0.188   9.78e-59
4 interaccion (Intercept)  6.75    1.54e-68
5 interaccion TV           0.0191  2.36e-27
6 interaccion radio        0.0289  1.40e- 3
7 interaccion TV:radio     0.00109 2.76e-51
```

Listing 5. *Resúmenes sobre los coeficientes.*

```
tibble(modelo = list(model.1, model.2)) ▷
  mutate(resultados = map(modelo, broom::glance)) ▷
  select(-modelo)▷
  unnest(resultados) ▷
  select(r.squared, sigma, AIC, deviance)
```

```
# A tibble: 2 × 4
  r.squared sigma   AIC deviance
      <dbl> <dbl> <dbl>    <dbl>
1     0.897 1.68   780.     557.
2     0.968 0.944  550.     174.
```

Listing 6. *Resúmenes globales de los modelos.*

El efecto de incrementar el presupuesto en un canal de ventas puede aumentar la efectividad de otro.

4.3. Jerarquías

¿Qué pasa cuando un valor-p de una interacción es pequeño, pero de los términos individuales no?

4.4. Interacciones y modelos múltiples

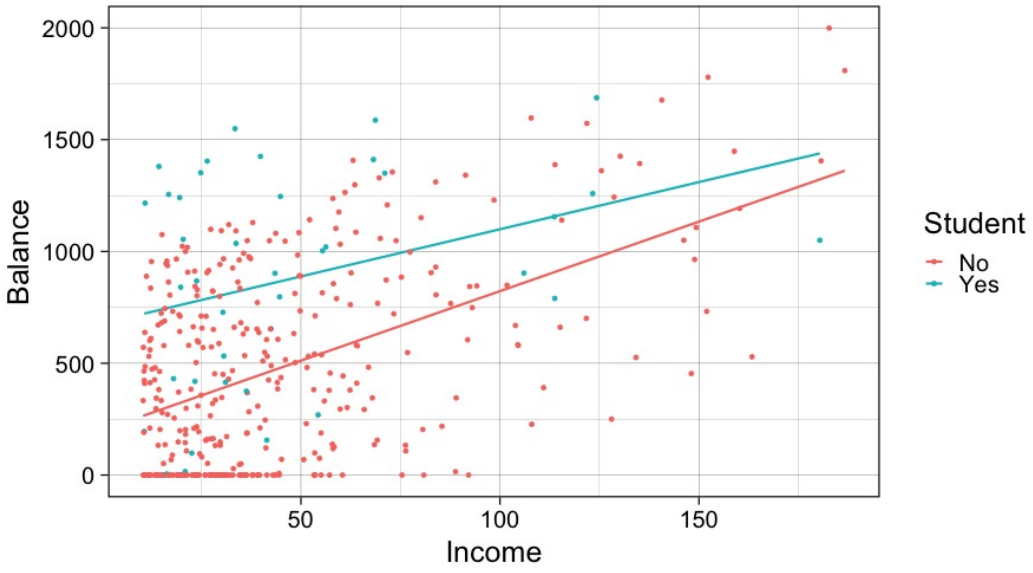


FIGURA 3. *Ajuste con interacción cualitativa y cuantitativa.*

4.5. Problemas con supuestos.

- No hay una relación lineal.
- Los errores están correlacionados.
- No hay varianza constante.
- Valores atípicos.
- Multicolinealidad.
- Puntos ancla.

5. GENERALIZACIONES

- Problemas de clasificación (siguiente).
- No-linealidad.
- Interacciones.
- Regularización.

6. APLICACIÓN: PREDICCIÓN DE ASISTENCIA EN PARTIDOS DE FOOTBALL

Tomado de esta liga y es un buen ejemplo de lo que se puede empezar a hacer con tidymodels [2].

```
1 library(tidyverse)
2 base_url ← "https://raw.githubusercontent.com/rfordatascience/tidytuesday/
      master/data/2020/2020-02-04/"
3
4 attendance ← read_csv(paste(base_url, "attendance.csv", sep = ""),
```

```
                        progress = FALSE, show_col_types = FALSE)
standings ← read_csv(paste(base_url, "standings.csv", sep = ""),
                     progress = FALSE, show_col_types = FALSE)

attendance_joined ← attendance ▷
  left_join(standings, by = c("year", "team_name", "team"))
attendance_joined
```

```
# A tibble: 10,846 × 20
   team     team_...¹n  year  total   home   away  week ...²weekl  wins  loss
      ...³point
   <chr>    <chr>      <dbl>  <dbl>  <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>   <
      dbl>
 1 Arizona ...Cardina  2000 893926 387475 506451     1   77434     3    13
        210
 2 Arizona ...Cardina  2000 893926 387475 506451     2   66009     3    13
        210
 3 Arizona ...Cardina  2000 893926 387475 506451     3      NA     3    13
        210
 4 Arizona ...Cardina  2000 893926 387475 506451     4   71801     3    13
        210
 5 Arizona ...Cardina  2000 893926 387475 506451     5   66985     3    13
        210
 6 Arizona ...Cardina  2000 893926 387475 506451     6   44296     3    13
        210
 7 Arizona ...Cardina  2000 893926 387475 506451     7   38293     3    13
        210
 8 Arizona ...Cardina  2000 893926 387475 506451     8   62981     3    13
        210
 9 Arizona ...Cardina  2000 893926 387475 506451     9   35286     3    13
        210
10 Arizona ...Cardina  2000 893926 387475 506451    10   52244     3    13
       210
# ... with 10,836 more rows, 9 more variables: points_against <dbl>,
#   points_differential <dbl>, margin_of_victory <dbl>,
#   strength_of_schedule <dbl>, simple_rating <dbl>, offensive_ranking <dbl>,
#   defensive_ranking <dbl>, playoffs <chr>, sb_winner <chr>, and abbreviated
#   variable names ¹team_name, ²weekly_attendance, ³points_for
# Use 'print(n = ...)' to see more rows, and 'colnames()' to see all variable
    names
```

```
  attendance_df ← attendance_joined ▷
  filter(!is.na(weekly_attendance)) ▷
  select(
    weekly_attendance, team_name, year, week,
    margin_of_victory, strength_of_schedule, playoffs
  )

attendance_df
```

```
# A tibble: 10,208 × 7
   weekly_attendance team_name  year  week margin_of_victory strength_...¹
      ...²playo
               <dbl> <chr>     <dbl> <dbl>             <dbl>        <dbl> <chr>
 1             77434 Cardinals  2000     1             -14.6         -0.7 No
    ...Pla
```

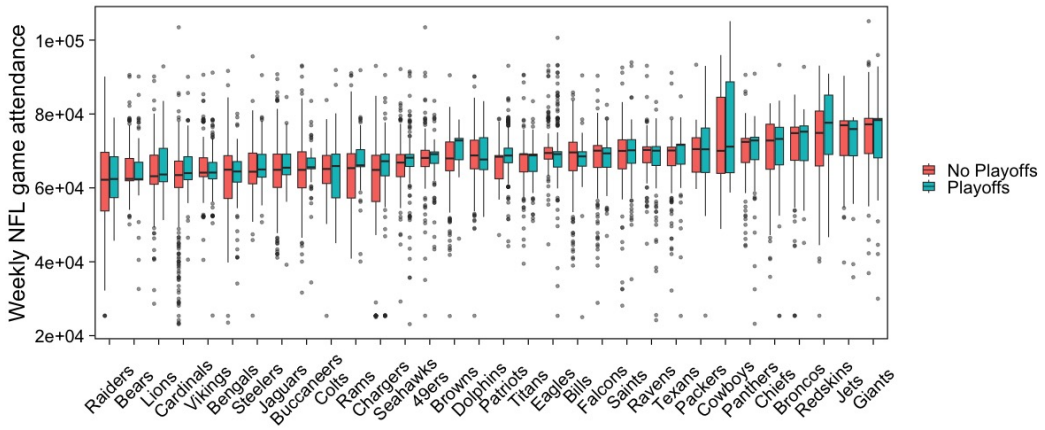


FIGURA 4. *Asistencia en estadios por equipo.*

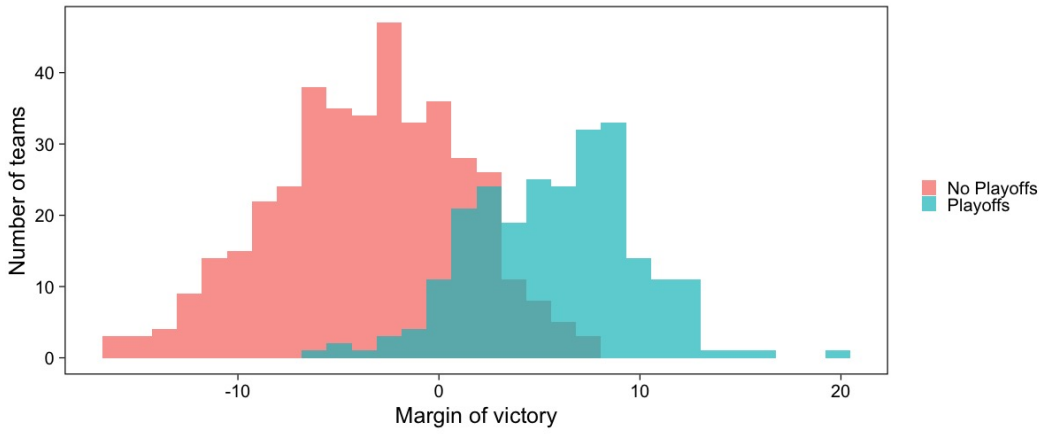


FIGURA 5. *Diferencia de puntos en partidos ganados.*

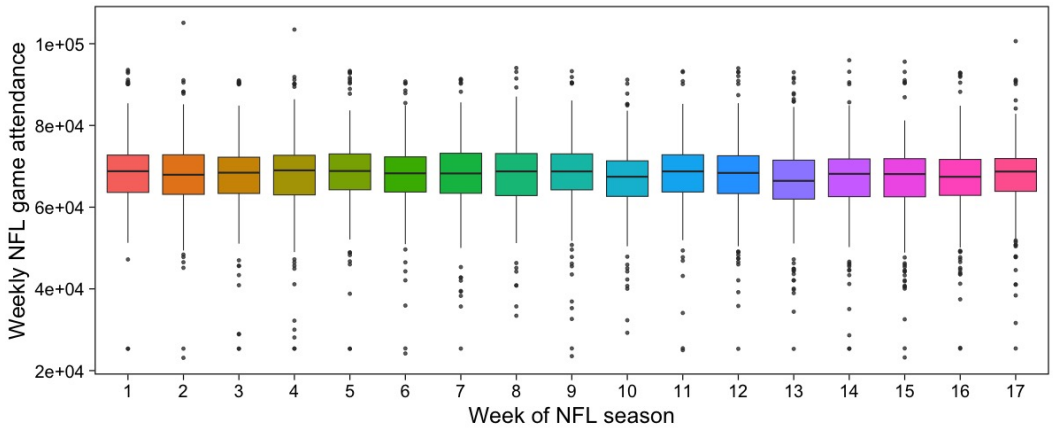


FIGURA 6. *Asistencia a lo largo de la temporada.*

```
 2             66009 Cardinals  2000     2             -14.6         -0.7 No
    ...Pla
 3             71801 Cardinals  2000     4             -14.6         -0.7 No
    ...Pla
 4             66985 Cardinals  2000     5             -14.6         -0.7 No
    ...Pla
 5             44296 Cardinals  2000     6             -14.6         -0.7 No
    ...Pla
 6             38293 Cardinals  2000     7             -14.6         -0.7 No
    ...Pla
 7             62981 Cardinals  2000     8             -14.6         -0.7 No
    ...Pla
 8             35286 Cardinals  2000     9             -14.6         -0.7 No
    ...Pla
 9             52244 Cardinals  2000    10             -14.6         -0.7 No
    ...Pla
10             64223 Cardinals  2000    11             -14.6         -0.7 No
   ...Pla
# ... with 10,198 more rows, and abbreviated variable names
# ¹  strength_of_schedule, ²playoffs
#  Use `print(n = ...)` to see more rows
```

```
library(tidymodels)

set.seed(108727)
attendance_split ← attendance_df ▷
  initial_split(strata = playoffs)

nfl_train ← training(attendance_split)
nfl_test ← testing(attendance_split)
```

```
lm_spec ← linear_reg() ▷
  set_engine(engine = "lm")

lm_spec
```

```
Linear Regression Model Specification (regression)

Computational engine: lm
```

```
lm_fit ← lm_spec ▷
  fit(weekly_attendance ~ .,
      data = nfl_train
      )

lm_fit ▷ broom::tidy()
```

```
# A tibble: 37 × 5
   term                estimate std.error statistic  p.value
   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)         -79125.      33598.    -2.36   1.85e- 2
 2 team_nameBears       -2613.        767.    -3.41   6.64e- 4
 3 team_nameBengals     -4757.        771.    -6.17   7.08e-10
```

```
 4 team_nameBills          106.       766.     0.138  8.90e- 1
 5 team_nameBroncos       2889.       775.     3.73   1.93e- 4
 6 team_nameBrowns         -21.6      775.    -0.0278 9.78e- 1
 7 team_nameBuccaneers   -2796.       752.    -3.72   2.04e- 4
 8 team_nameCardinals    -5905.       767.    -7.70   1.53e-14
 9 team_nameChargers     -5098.       774.    -6.59   4.76e-11
10 team_nameChiefs        1802.       763.     2.36   1.83e- 2
# ... with 27 more rows
#  Use 'print(n = ...)' to see more rows
```

```
lm_fit ▷ broom::tidy(conf.int = TRUE)
```

```
# A tibble: 37 × 7
   term                estimate std.error statistic  p.value conf.low conf.
      high
   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl
      >
 1 (Intercept)          -79125.     33598.    -2.36   1.85e- 2 -144985.
    -13264.
 2 team_nameBears        -2613.       767.    -3.41   6.64e- 4   -4117.
    -1109.
 3 team_nameBengals      -4757.       771.    -6.17   7.08e-10   -6268.
    -3246.
 4 team_nameBills          106.       766.     0.138  8.90e- 1   -1396.
     1608.
 5 team_nameBroncos       2889.       775.     3.73   1.93e- 4    1371.
     4408.
 6 team_nameBrowns         -21.6      775.    -0.0278 9.78e- 1   -1541.
     1498.
 7 team_nameBuccaneers   -2796.       752.    -3.72   2.04e- 4   -4271.
    -1321.
 8 team_nameCardinals    -5905.       767.    -7.70   1.53e-14   -7408.
    -4401.
 9 team_nameChargers     -5098.       774.    -6.59   4.76e-11   -6615.
    -3581.
10 team_nameChiefs        1802.       763.     2.36   1.83e- 2     306.
    3298.
# ... with 27 more rows
#  Use 'print(n = ...)' to see more rows
```

```
lm_fit ▷ broom::glance()
```

```
# A tibble: 1 × 12
  r.squared adj.r...¹. sigma ...²stati   p.value    df  logLik    AIC    BIC
     ...³devia
      <dbl>     <dbl> <dbl>    <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <
         dbl>
1     0.152     0.148 8379.     38.1 7.30e-242    36 -80005. 1.60e5 1.60e5 5.35
    e11
# ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names ¹adj.r.squared, ²statistic, ³deviance
#  Use 'colnames()' to see all variable names
```

```
results_train ← lm_fit ▷
  predict(new_data = nfl_train) ▷
  mutate(truth = nfl_train$weekly_attendance,
         model = "lm")
results_train ▷
  rmse(truth = truth, estimate = .pred)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       8358.
```

```
results_test ← lm_fit ▷
  predict(new_data = nfl_test) ▷
  mutate(truth = nfl_test$weekly_attendance,
         model = "lm")
results_test ▷
  rmse(truth = truth, estimate = .pred)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       8192.
```

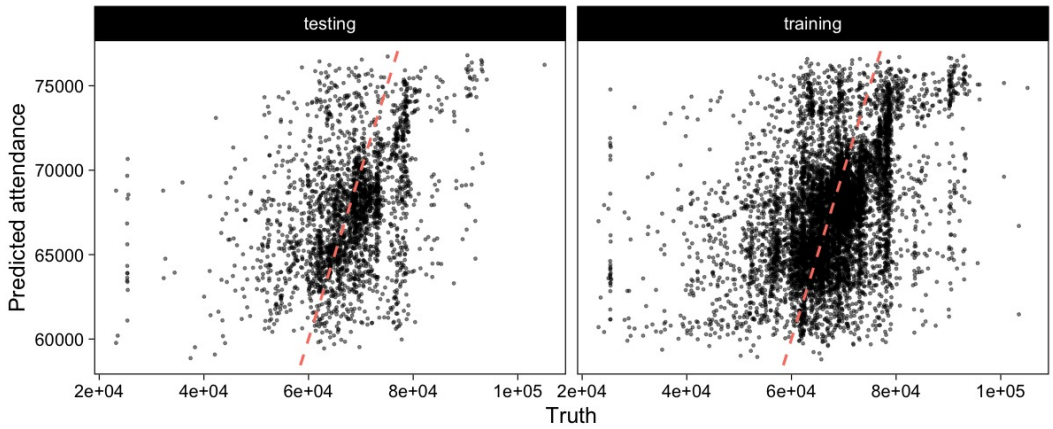


FIGURA 7. *Comparativo de predicciones contra valores reales.*

REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Texts in Statistics. Springer US, New York, NY, 2021. 1

[2] M. Kuhn and J. Silge. *Tidy Modeling with R.* .º'Reilly Media, Inc.", 2022. 8