EST-25134: Aprendizaje Estadístico

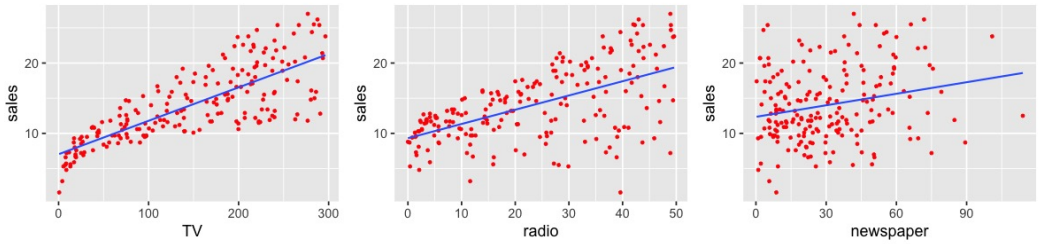
Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Regresión lineal.
Objetivo. Repasaremos los conceptos de regresión lineal desde un punto de vista de inferencia. Haremos conexiones interesantes con conceptos clave en probabilidad y teoria de la información. Veremos métricas de desempeño para modelos 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.

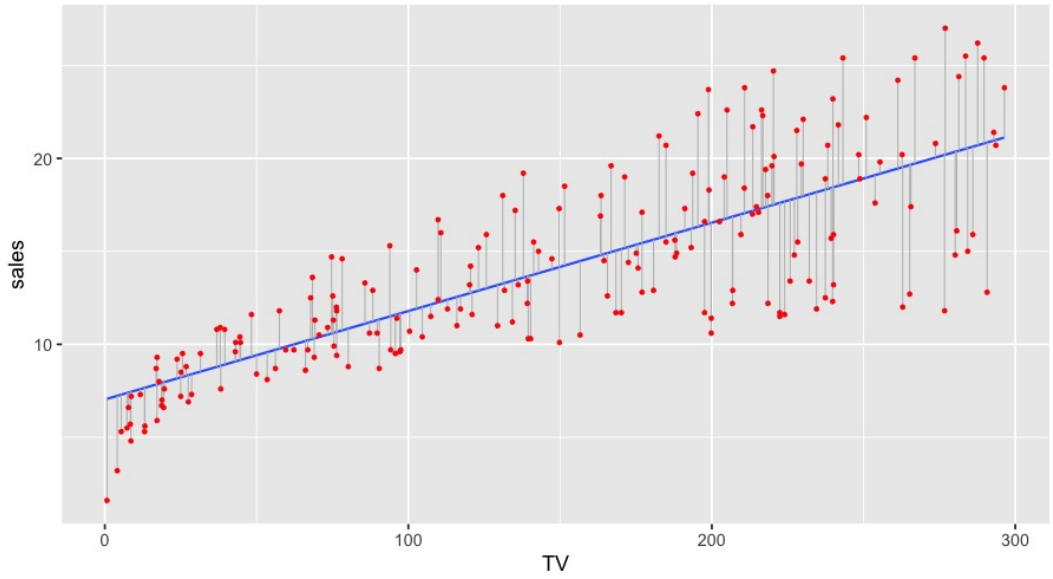


FIGURA 1. *Ajuste y residuales a la recta de mínimos cuadrados.*

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: 3.3 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() ▷
  as.data.frame()
```

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

```
         term estimate std.error statistic p.value
1 (Intercept)    7.033    0.4578        15 1.4e-35
2          TV    0.048    0.0027        18 1.5e-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.

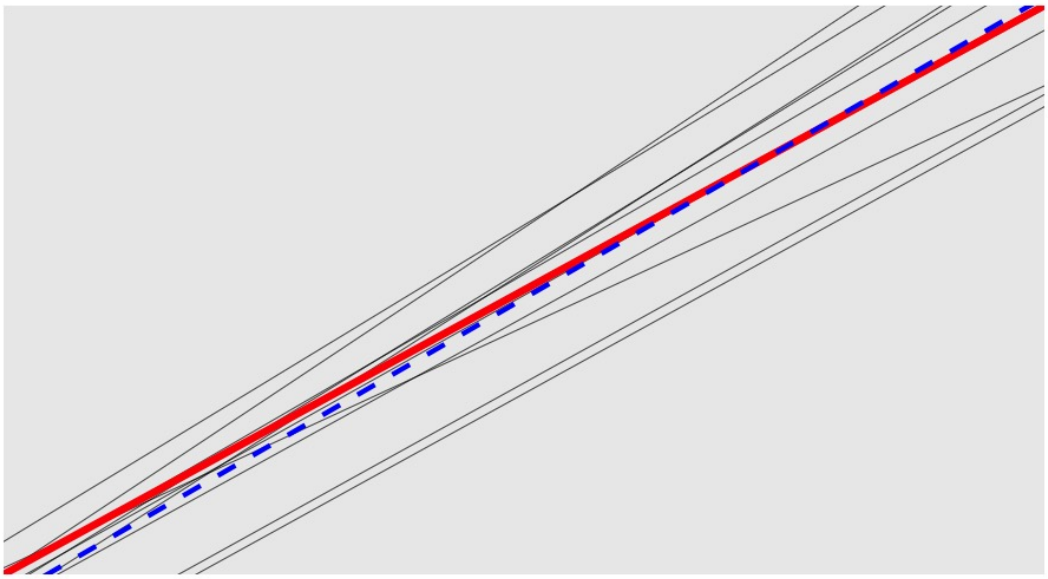


FIGURA 2. *Simulación de ajuste (variación en datos).*

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() ▷
  as.data.frame()
```

```
         term estimate std.error statistic p.value
1 (Intercept)    2.939    0.3119      9.42 1.3e-17
2          TV    0.046    0.0014     32.81 1.5e-81
3       radio    0.189    0.0086     21.89 1.5e-54
4   newspaper   -0.001    0.0059     -0.18 8.6e-01
```

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) ▷
  as.data.frame()
```

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

```
  statistic p.value df df.residual
1       570 1.6e-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)
```

```
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) ▷
  as.data.frame()
```

```
         tipo        term estimate p.value
1      lineal (Intercept)   2.9211 4.6e-19
2      lineal          TV   0.0458 5.4e-82
3      lineal       radio   0.1880 9.8e-59
4 interaccion (Intercept)   6.7502 1.5e-68
5 interaccion          TV   0.0191 2.4e-27
6 interaccion       radio   0.0289 1.4e-03
7 interaccion    TV:radio   0.0011 2.8e-51
```

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

```
  r.squared sigma AIC deviance
1      0.90  1.68 780      557
2      0.97  0.94 550      174
```

> 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

4.5. Problemas con supuestos.

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

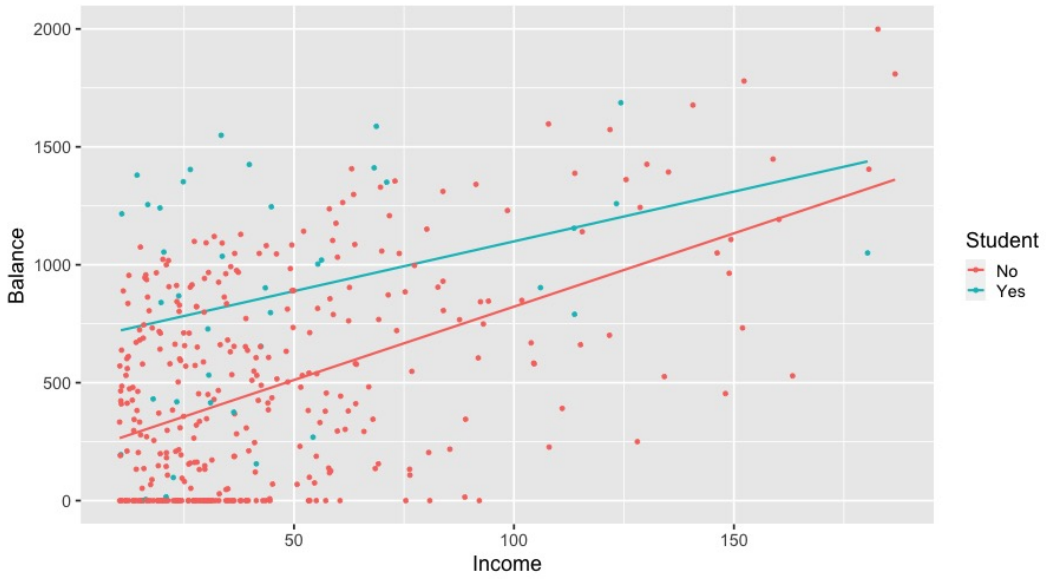


FIGURA 3. *Ajuste con interacción cualitativa y cuantitativa.*

5. GENERALIZACIONES

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

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. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 1