Tema 2 - Regresión lineal

Aprendizaje estadístico

Lorena Romero Mateo - 77857300T

Email: lorena.romerom@um.es

Índice

1. Regresión lineal

Es un problema de álgebra lineal, queremos minimizar la distancia euclídea.

La regresión lineal es un enfoque simple para el aprendizaje supervisado. Asume que la dependencia de Y en $X_1, X_2, ...X_p$ es lineal.

- ¡Las funciones de regresión verdaderas nunca son lineales!
- Aunque pueda parecer demasiado simplista, la regresión lineal es extremadamente útil tanto conceptual como prácticamente.

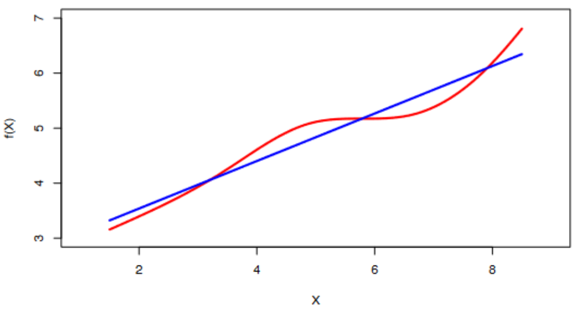

Figura 1: Regresión lineal

1.1. Regresión lineal simple usando un solo predictor X

Asumimos un modelo

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

donde β_0 y β_1 son dos constantes desconocidas que representan la intersección y la pendiente, también conocidas como coeficientes o parámetros, y ϵ es el término de error.

- Dadas algunas estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ para los coeficientes del modelo, predecimos ventas futuras usando

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

  donde $\hat{y}$ indica una predicción de Y sobre la base de $X = x$. El símbolo de sombrero denota un valor estimado. Nunca vamos a poder calcular los valores reales.

1.2. Estimación de los parametros por mínimos cuadrados

Siempre cae en el test.
Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y basada en el valor i-ésimo de X. Entonces $e_i = y_i - \hat{y}_i$ representa el residuo i-ésimo.

- Definimos la suma de los residuos al cuadrado (RSS) como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente como

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

RSS = Suma residual de cuadrados

Es la norma al cuadrado, distancia euclídea entre el vector a predecir y el modelo.

$$RSS = ||y - f(x)||^2$$

Propiedad: la media de los residuos es igual a 0.

Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y basada en el valor i-ésimo de X. Entonces $e_i = y_i - \hat{y}_i$ representa el residuo i-ésimo.

- Definimos la suma de los residuos al cuadrado (RSS) como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente como

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- El enfoque de mínimos cuadrados elige $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar el RSS. Los valores que minimizan pueden demostrarse como

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

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

donde $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ y $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ son las medias muestrales.

Estamos calculando una proyección ortogonal.

La recta de regresión es

$$\frac{y - \bar{y}}{\sigma y} = \rho \frac{x - \bar{x}}{\sigma x}$$

El índice de correlación es

$$\rho = \frac{\sum_{i=1}^n (x - x_i)(y - y_i)}{||x - \bar{x}|| \cdot ||y - \bar{y}||} \in [-1, 1]$$

entre -1 y 1 porque es un coseno

1.3. Regresión lineal para los datos de publicidad

Consideremos los datos de publicidad mostrados en la siguiente diapositiva. Preguntas que podríamos hacernos:

- ¿Existe una relación entre el presupuesto de publicidad y las ventas?
- ¿Qué tan fuerte es la relación entre el presupuesto de publicidad y las ventas?
- ¿Qué medios contribuyen a las ventas?
- ¿Qué tan precisamente podemos predecir las ventas futuras?
- ¿Es la relación lineal?
- ¿Existe sinergia entre los medios publicitarios?

Práctica de 1º semana responde a estas preguntas.

1.4. Error estándar de los estimadores de los parámetros

El error estándar de un estimador refleja cómo varía bajo muestreo repetido. Tenemos

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

donde $\sigma^2 = \mathrm{Var}(\epsilon)$. que se estima con la varianza residual Estas expresiones se pueden considerar como la varianza de los estimadores. La matriz de varianza-covarianza de los estimadores de los parámetros es:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

El error estándar de un estimador refleja cómo varía bajo muestreo repetido. Tenemos

$$SE(\hat{\beta}_1)^2 = \frac{sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

donde $\sigma^2 = \mathrm{Var}(\epsilon)$.

- Estos errores estándar se pueden usar para calcular intervalos de confianza. Un intervalo de confianza del 95 % se define como un rango de valores tal que, con un 95 % de probabilidad, el rango contendrá el valor verdadero desconocido del parámetro. Tiene la forma $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$.

Si tenemos una variable continua que sigue la distribución normal, la probabilidad de encontrarla en un intervalo es del 95 %.

$$X \sim N(\mu, \sigma^2) \Rightarrow P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}95$$

1.5. Intervalos de confianza

Es decir, hay aproximadamente un 95

$$\left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\right]$$

contenga el valor verdadero de β_1 (bajo un escenario donde obtuvimos muestras repetidas como la muestra presente). Para los datos de publicidad, el intervalo de confianza del 95

$$[0{,}042, 0{,}053]$$

1.6. Pruebas de hipótesis

Los errores estándar también se pueden usar para realizar pruebas de hipótesis sobre los coeficientes. La prueba de hipótesis más común implica probar la hipótesis nula de

$$H_0 : \text{No hay relación entre } X \text{ e } Y$$

versus la hipótesis alternativa

$$H_A : \text{Hay alguna relación entre } X \text{ e } Y.$$

- Matemáticamente, esto corresponde a probar

  $$H_0 : \beta_1 = 0$$

  versus

  $$H_A : \beta_1 \neq 0,$$

  ya que si $\beta_1 = 0$ entonces el modelo se reduce a $Y = \beta_0 + \epsilon$, y X no está asociado con Y.

Para probar la hipótesis nula, calculamos una estadística t, dada por

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)},$$

- Esta tendrá una distribución t con $n - 2$ grados de libertad, asumiendo que $\beta_1 = 0$.
- Usando software estadístico, es fácil calcular la probabilidad de observar cualquier valor igual a $|t|$ o mayor. Llamamos a esta probabilidad el valor p.

1.7. Evaluación de la Precisión General del Modelo

Calculamos el Error Estándar Residual

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

donde la suma de los residuos al cuadrado es

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

RSE = Error estándar residual
RSS/(n-2) = varianza residual
MSE = mean squared error

En este modelo, la variabilidad total de Y (TSS) se descompone en la variabilidad explicada por el modelo (VE) y la residual (RSS). El coeficiente de determinación, R^2, es la proporción de la variabilidad total explicada por el modelo de regresión:

$$R^2 = \frac{VE}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Calculamos el Error Estándar Residual

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

donde la suma de los residuos al cuadrado es

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

R-cuadrado o fracción de varianza explicada es

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

donde $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ es la suma total de cuadrados.

En este modelo, la variabilidad total de Y (TSS) se descompone en la variabilidad explicada por el modelo (VE) y la residual (RSS). El coeficiente de determinación, R^2, es la proporción de la variabilidad total explicada por el modelo de regresión:

$$R^2 = \frac{VE}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Se puede demostrar que en este contexto de regresión lineal simple, $R^2 = r^2$, donde r es la correlación entre X e Y:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

2. Regresión lineal múltiple

Aquí nuestro modelo es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Interpretamos β_j como el efecto promedio en Y de un aumento de una unidad en X_j, manteniendo todos los demás predictores fijos. En el ejemplo de publicidad, el modelo se convierte en

$$\text{ventas} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{periódico} + \epsilon$$

2.1. Interpretando coeficientes de regresión

El escenario ideal es cuando los predictores no están correlacionados — un diseño equilibrado:

- Cada coeficiente puede ser estimado y probado por separado.
- Interpretaciones como un cambio de una unidad en X_j está asociado con "un cambio de β_j en Y, mientras que todas las demás variables permanecen fijas", son posibles.

Las correlaciones entre los predictores causan problemas:

- La varianza de todos los coeficientes tiende a aumentar, a veces de manera dramática.
- Las interpretaciones se vuelven peligrosas — cuando X_j cambia, todo lo demás cambia.

Las **afirmaciones de causalidad** deben evitarse para datos observacionales. Evitar falsas conclusiones.

(Anteriormente vimos que existe mucho sesgo y poca varianza, si tengo varios predictores entonces puedo tener más varianza).

Dadas las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, podemos hacer predicciones usando la fórmula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

- Estimamos $\beta_0, \beta_1, \ldots, \beta_p$ como los valores que minimizan la suma de los residuos al cuadrado

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2.$$

Esto se hace usando software estadístico estándar. Los valores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ que minimizan el RSS son las estimaciones de los coeficientes de regresión por mínimos cuadrados múltiples.

2.2. Error estándar residual y R-cuadrado

- El error estándar residual, RSE, proporciona una medida absoluta de la falta de ajuste del modelo, pero depende de las unidades de Y.
- En el modelo lineal múltiple, el RSE se estima como

$$RSE = \sqrt{\frac{RSS}{n-p-1}}$$

- Para calcular R^2, usamos la misma fórmula,

$$R^2 = 1 - \frac{RSS}{TSS}$$

2.3. Algunas cuestiones importantes

1. ¿Es al menos uno de los predictores $X_1, X_2, \ldots, X_p$ útil para predecir la respuesta?
2. ¿Todos los predictores ayudan a explicar Y, o solo un subconjunto de los predictores es útil?
3. ¿Qué tan bien se ajusta el modelo a los datos?
4. Dado un conjunto de valores de predictores, ¿qué valor de respuesta deberíamos predecir y qué tan precisa es nuestra predicción?

2.4. ¿Es al menos un predictor útil?

Para la primera pregunta, podemos usar la estadística F

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

Si todos los parámetros son nulos, F se debe aproximar a 1. Si algún parámetro no es nulo, F debería ser mayor que 1.

Quantity	**Value**
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

2.5. Decidiendo variables importantes

El enfoque más directo se llama regresión de todos los subconjuntos o mejores subconjuntos: calculamos el ajuste de mínimos cuadrados para todos los subconjuntos posibles y luego elegimos entre ellos basándonos en algún criterio que equilibre el error de entrenamiento con el tamaño del modelo.

Sin embargo, a menudo no podemos examinar todos los modelos posibles, ya que hay 2^p de ellos; por ejemplo, cuando $p = 40$ hay más de mil millones de modelos.

En su lugar, necesitamos un enfoque automatizado que busque a través de un subconjunto de ellos. A continuación, discutimos dos enfoques comúnmente utilizados.

2.6. Selección hacia adelante

- Comenzar con el modelo nulo — un modelo que contiene una intersección pero no predictores.
- Ajustar p regresiones lineales simples y añadir al modelo nulo la variable que resulta en el menor RSS.
- Añadir a ese modelo la variable que resulta en el menor RSS entre todos los modelos de dos variables.
- Continuar hasta que se cumpla alguna regla de parada, por ejemplo, cuando todas las variables restantes tengan un valor p por encima de algún umbral.

2.7. Selección hacia atrás

- Comenzar con todas las variables en el modelo.
- Eliminar la variable con el valor p más grande, es decir, la variable que es menos estadísticamente significativa.
- Ajustar el nuevo modelo con $p - 1$ variables, y eliminar la variable con el valor p más grande.
- Continuar hasta que se alcance una regla de parada. Por ejemplo, podemos detenernos cuando todas las variables restantes tengan un valor p significativo definido por algún umbral de significancia.

2.8. Otras consideraciones en el modelo de regresión

2.8.1. Predictores cualitativos

- Algunos predictores no son cuantitativos, sino cualitativos, tomando un conjunto discreto de valores.
- Estos también se llaman predictores categóricos o variables de factor.

- Véase, por ejemplo, la matriz de dispersión de los datos de tarjetas de crédito en la siguiente diapositiva.

Además de las 7 variables cuantitativas mostradas, hay cuatro variables cualitativas: género, estudiante (estado de estudiante), estado (estado civil) y etnicidad (caucásico, afroamericano (AA) o asiático).

Ejemplo: investigar las diferencias en el saldo de la tarjeta de crédito entre hombres y mujeres, ignorando las otras variables. Creamos una nueva variable

$$x_i = \begin{cases} 1 & \text{si la persona } i \text{ es mujer} \\ 0 & \text{si la persona } i \text{ es hombre} \end{cases}$$

Modelo resultante:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es mujer} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es hombre} \end{cases}$$

Interpretación?

3. Extensiones del Modelo Lineal

Eliminando la suposición aditiva: interacciones y no linealidad

3.1. Interacciones

- En nuestro análisis previo de los datos de Publicidad, asumimos que el efecto sobre las ventas de aumentar un medio publicitario es independiente de la cantidad gastada en los otros medios.
- Por ejemplo, el modelo lineal

  $$\widehat{\text{ventas}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{periódico}$$

  establece que el efecto promedio sobre las ventas de un aumento de una unidad en TV es siempre β_1, independientemente de la cantidad gastada en radio.
- Pero supongamos que gastar dinero en publicidad en radio realmente aumenta la efectividad de la publicidad en TV, de modo que el término de pendiente para TV debería aumentar a medida que aumenta la radio.
- En esta situación, dado un presupuesto fijo de $100,000, gastar la mitad en radio y la mitad en TV puede aumentar las ventas más que asignar la cantidad total a TV o a radio.
- En marketing, esto se conoce como un efecto de sinergia, y en estadística se refiere como un efecto de interacción.