Minería de Datos
IIC2433

Modelos de Regresión

Vicente Domínguez

¿Qué veremos esta clase?

- Modelos de Regresión
- Cómo predecir una variable numérica

Métodos de aprendizaje

- Aprendizaje supervisado
  - Necesita conocimiento previo del problema y el valor a predecir
  - Se pueden usar valores numéricos o etiquetas
- Aprendizaje no supervisado
  - No se necesita conocimiento previo
  - Modelos buscan patrones dentro del conjunto de datos

Métodos de aprendizaje

- Aprendizaje supervisado **(necesita etiquetas)**
  - Clasificación
  - **Regresión**

- Aprendizaje no supervisado **(no necesita etiquetas)**
  - Clustering
  - Reglas de asociación
  - etc

Regresiones Lineales

- Técnica estadística donde se trata de ajustar parámetros de una función lineal sobre un conjunto de datos.
- Se busca predecir el valor de una variable dependiente cuantitativa (predicha) utilizando variables independientes (predictores)
- Finalmente, queremos determinar cómo afecta nuestra variable independiente a la dependiente

$$\boldsymbol{Y} = \alpha + \beta \boldsymbol{X}$$

Volviendo a lo básico

- Dada una tabla con un conjunto de atributos numéricos $A_1, \ldots, A_n$
- Se busca predecir un atributo numérico B
- Asumimos que esta tabla representa una función $f : \mathbb{R}^n \to \mathbb{R}$ la cual buscamos aprender
- Y que dicha función, es una función lineal, es decir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Volviendo a lo básico

- Viéndolo desde otra perspectiva, estamos suponiendo que nuestra función es una **transformación lineal** que toma vectores en $\mathbb{R}^n$ y los transforma en un número en $\mathbb{R}$
- Se dice lineal porque se obtiene con una simple multiplicación entre 2 vectores. Sea $\bar{x} = (x_1, \ldots, x_n)$ y $\beta = (\beta_1, \ldots, \beta_n)$ entonces, la transformación se obtiene como:

$$y = \beta_0 + \beta^T \cdot \bar{x}$$

¿Por qué regresiones lineales?

- Se puede tener una noción de qué parámetros son importantes y cuáles no. Basta con saber que Beta tiene un valor más grande.
- Se puede obtener una función de forma analítica, lo cual no siempre es posible.

Regresiones Lineales

Condiciones y supuestos

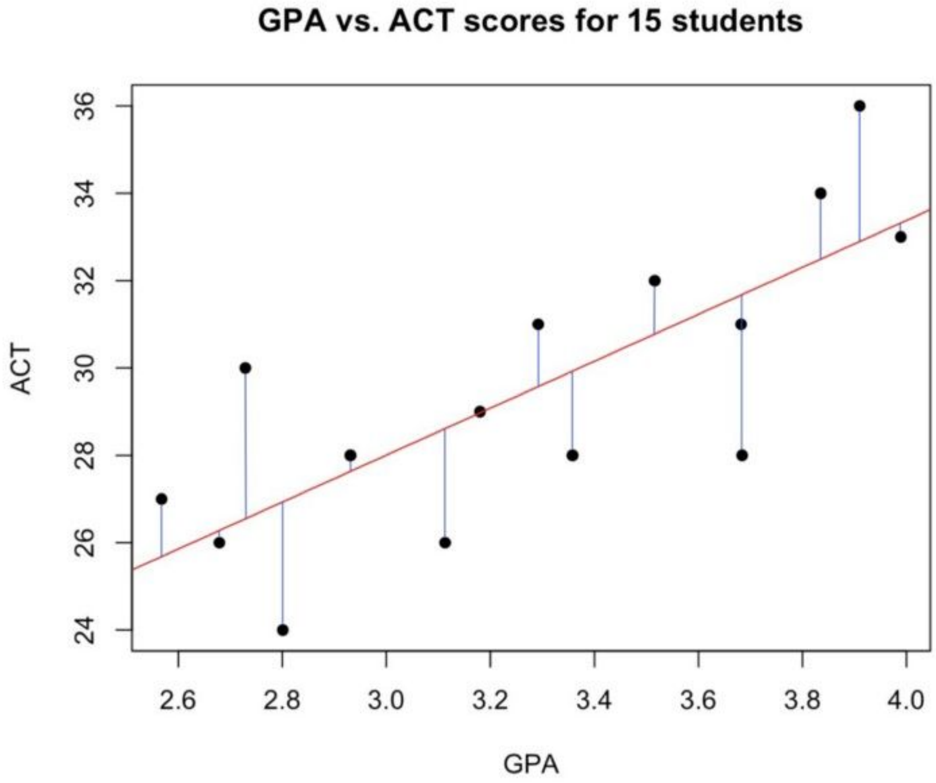
Para que un modelo de regresión lineal funcione correctamente, deben cumplirse ciertas condiciones.

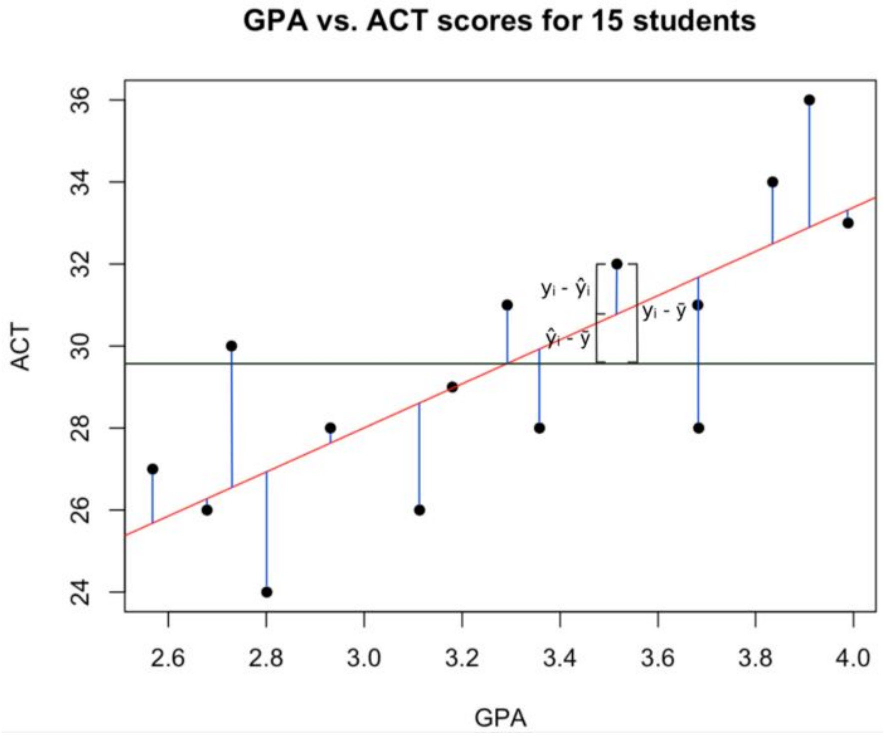
- Homocedasticidad
- Independencia
- Normalidad

Aparte también hay medidas que nos permiten evaluar que tan bien ajusta el modelo:

- R^2 o varianza explicada
- RSS o suma de errores residuales

Ordinary Least Squares





Fuente: https://towardsdatascience.com/how-least-squares-regression-estimates-are-actually-calculated-662d237a4d7e

Ordinary Least Squares

Busca minimizar el error cuadrático residual RSS

$$Y = \alpha + \beta X$$

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{\mathrm{Cov}[x, y]}{\mathrm{Var}[x]}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

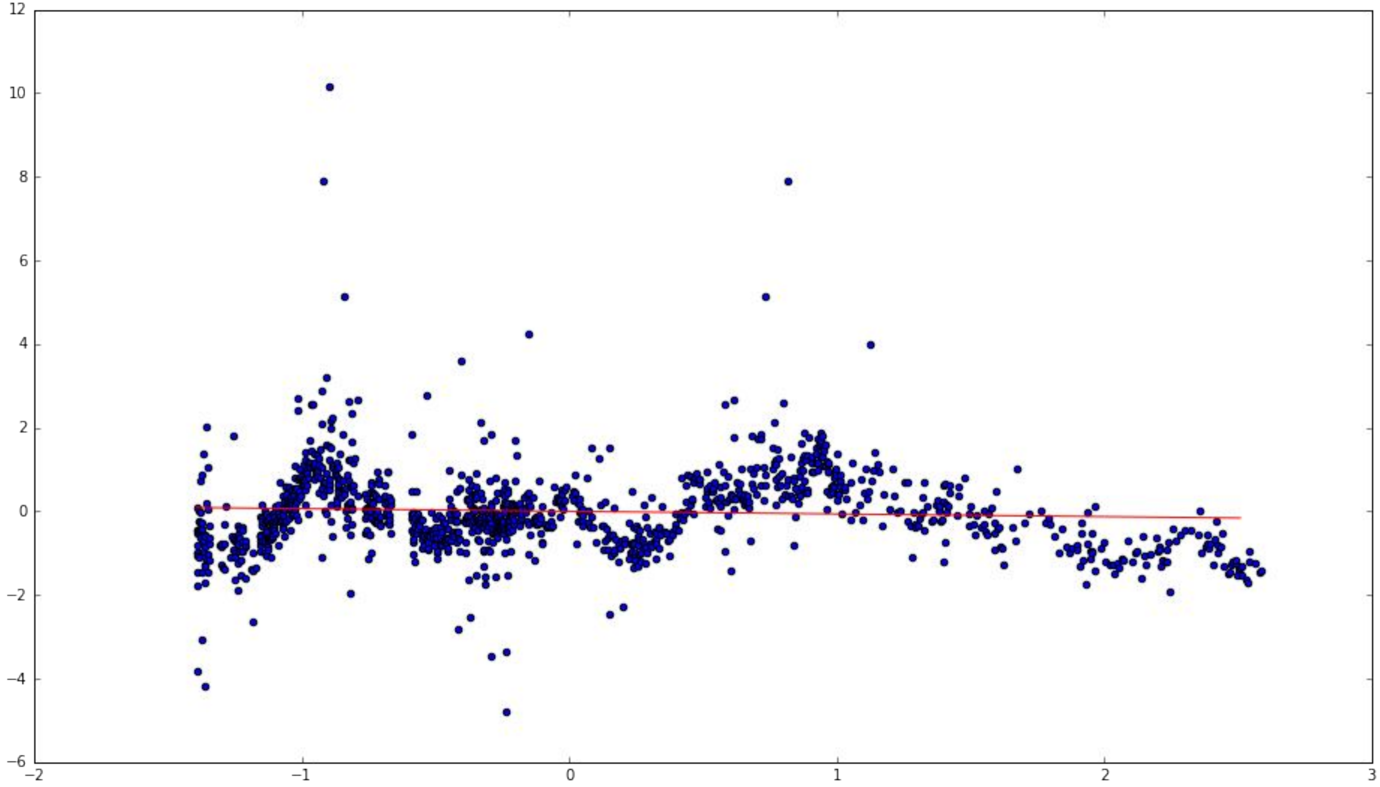
$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = \left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = 0$$

$$
\begin{cases}
-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0 \\
-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)x_i=0
\end{cases}
$$

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

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Modelo de regresión simple



Regresiones no lineales

- Siempre se puede hacer un cambio de *Kernel* sobre el conjunto de datos.
- Es decir, se puede hacer una transformación con combinaciones lineales, de funciones no lineales sobre el conjunto de datos

$$f(X) \to f(\Phi(X))$$

Donde $\Phi(X)$ es una transformación sobre X, por ejemplo

$$\Phi(X) \to [X, X^2, X^4]$$

Regresiones no lineales

- Ajustar una regresión lineal no significa **ajustar una recta**.
- Si ajustamos una línea sobre un espacio curvo nos debería ajustar una curva.
- En particular también podemos ajustar cualquier función no lineal como *Kernel*, por ejemplo una Gaussiana

$$f(X) = e^{-\frac{1}{2}\left(\frac{X-\mu_i}{\sigma}\right)^2}$$

Para un kernel de tamaño 3, la matemática es la siguiente

$$\Phi_i(X) = e^{-\frac{1}{2}(\frac{X-\mu_i}{\sigma})^2}$$

$$\Phi(X) = [\Phi_1(X), \Phi_2(X), \Phi_3(X)]$$

$$Y = [Y_1, Y_2, Y_3]$$

$$W = \Phi(X)/sum(\Phi(X)) = [w_1, w_2, w_2]$$

$$\hat{Y} = WY$$

Regresiones no lineales