Linear Regression

Lê Hồng Phương
Data Science Laboratory
Vietnam National University, Hanoi
<*phuonglh@hus.edu.vn*>

October 11, 2021

Linear Analysis

- **Phân tích hồi quy** nghiên cứu sự phụ thuộc của một *biến trả lời* (response variable) vào một hoặc nhiều *biến dự báo* (predictors).
  - y: biến trả lời
  - $(x_1, x_2, \ldots, x_D)$: các biến dự báo
- Đây là kĩ thuật rất cơ bản của ngành thống kê toán học. Để hiểu được nhiều mô hình phân loại, dự báo hiện đại, ta cần hiểu rõ các phương pháp phân tích hồi quy cổ điển.
- Có 2 dạng phân tích hồi quy:
  - hồi quy tuyến tính
  - hồi quy phi tuyến
- Phân tính hồi quy đơn: $y \in \mathbb{R}$; phân tích hồi quy bội $y \in \mathbb{R}^n, n > 1$.

Linear Regression

Hàm dự báo $h_\theta(\mathrm{x})$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(\mathrm{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_D x_D.$$

Nếu bổ sung thêm đặc trưng cố định $x_0 \equiv 1$ thì ta có thể biểu diễn h dưới dạng

$$h_\theta(\mathrm{x}) = \sum_{j=0}^{D} \theta_j x_j = \theta^\top \mathrm{x}\,.$$

Linear Regression

Tập dữ liệu huấn luyện:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$$

Để ước lượng tham số $\theta \in \mathbb{R}^{D+1}$, ta cực tiểu hóa sai số của mô hình trên tập dữ liệu huấn luyện:

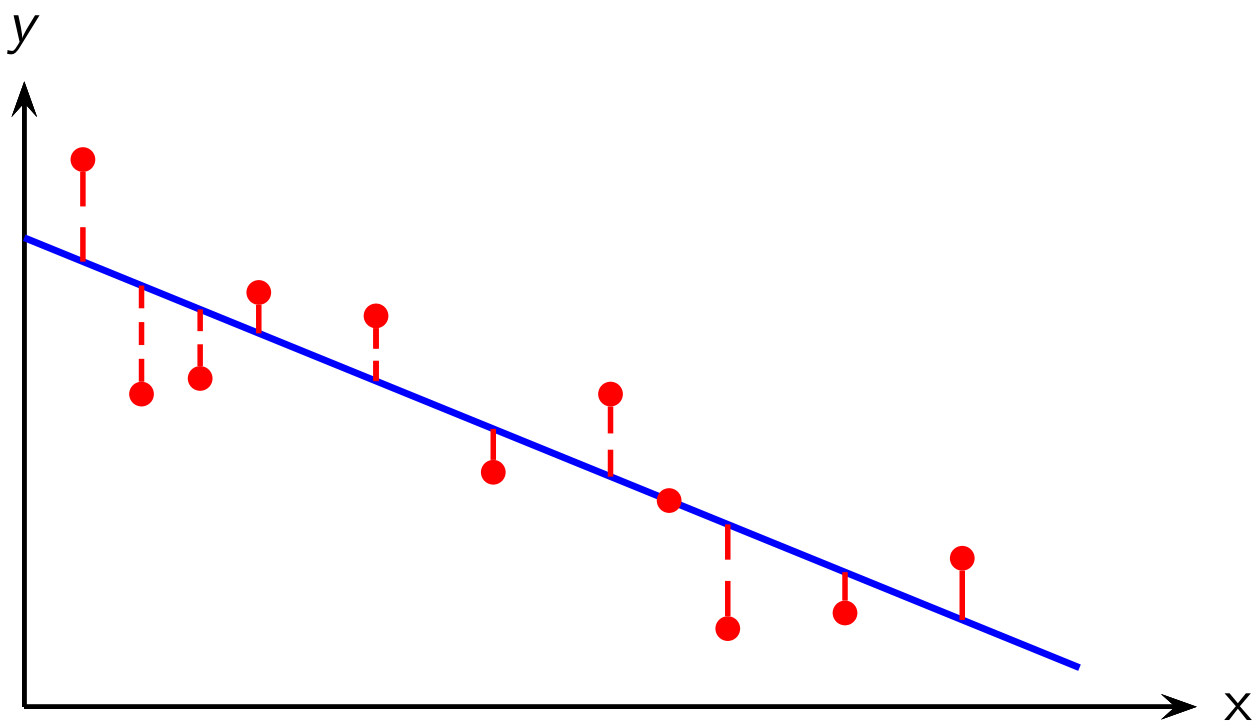
$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2.$$

Phương pháp này được gọi là **bình phương tối thiểu** (OLS – *Ordinary Least Squares*)

$$J(\theta) \to \min.$$

Linear Regression

Các điểm dữ liệu (x, y) được dự báo bởi một siêu phẳng trong không gian nhiều chiều $h_\theta(\mathrm{x}) = \theta^\top \mathrm{x}$.



Linear Regression

Để dự đoán giá trị y cho mỗi đối tượng x, ta dùng hàm $h_\theta(\mathrm{x})$ với sai khác là một *nhiễu ngẫu nhiên* ϵ:

$$y = h_\theta(\mathrm{x}) + \epsilon.$$

Giả sử ϵ tuân theo phân phối chuẩn một chiều

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Hàm mật độ của ϵ:

$$P(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Từ đó ta có

$$P(y|\mathrm{x};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - h_\theta(\mathrm{x}))^2}{2\sigma^2}\right).$$

Linear Regression

- Log-hợp lí của dữ liệu là:

$$L(\theta) = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - h_\theta(\mathrm{x}_i))^2}{2\sigma^2} \right) \right]$$

$$= N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - h_\theta(\mathrm{x}_i))^2.$$

- Phương pháp hợp lí cực đại cực tiểu hàm mục tiêu

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} [h_\theta(\mathrm{x}_i) - y_i]^2.$$

- Đây chính là hàm sai số trong phương pháp bình phương tối thiểu.

Linear Regression

Như vậy:

- Nếu giả định nhiễu ngẫu nhiên tuân theo phân phối chuẩn thì phương pháp ước lượng hợp lí cực đại dẫn tới phương pháp hồi quy bình phương tối thiểu.
- Ta thấy θ không phụ thuộc vào phương sai σ^2, ngay cả nếu không biết σ^2 thì ta vẫn ước lượng được θ.

Normal Equation

Ta có thể tìm được nghiệm đúng dạng giải tích của θ bằng phương trình chuẩn.

- Ta có **ma trận thiết kế** X cỡ $N \times (D+1)$

$$\mathrm{X} = \begin{pmatrix} (\mathrm{x}_1)^\top \\ (\mathrm{x}_2)^\top \\ \vdots \\ (\mathrm{x}_N)^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND} \end{pmatrix}.$$

- Đặt y là véc-tơ cột chứa tất cả các giá trị của biến dự báo:

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Normal Equation

Viết lại hàm mục tiêu dưới dạng ma trận:

$$J(\theta) = \frac{1}{2}(\mathrm{X}\theta - \mathrm{y})^\top(\mathrm{X}\theta - \mathrm{y}).$$

Đạo hàm của $J(\theta)$ ứng với θ là

$$\nabla J(\theta) = \mathrm{X}^\top\mathrm{X}\theta - \mathrm{X}^\top\mathrm{y}.$$

Để cực tiểu hoá J, ta tìm θ thoả mãn $\nabla J(\theta) = 0$, từ đó có nghiệm đúng cho θ là

$$\boxed{\hat{\theta} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y}.}$$

Phương trình trên được gọi là **phương trình chuẩn**.

Normal Equation

- Nếu tồn tại nghịch đảo của ma trận $X^\top X$ thì ta mới tìm được nghiệm duy nhất θ theo phương trình chuẩn.
- Nếu nghịch đảo không tồn tại, tức là $X^\top X$ không đủ hạng thì ước lượng hồi quy là không duy nhất.
  - Tồn tại một phụ thuộc tuyến tính giữa các cột của X.
  - Ta cần tìm cách giảm số chiều của x bằng việc loại bỏ các đặc trưng phụ thuộc sao cho ma trận thiết kế là đủ hạng.

Some Properties

Sử dụng công thức $\text{var}(a + Ay) = A\,\text{var}(y)A^\top$ với a, y là các véc-tơ và A là ma trận hằng số, ta có phương sai của $\hat{\theta}$ là

$$\begin{aligned}
\text{var}(\hat{\theta}|X) &= \text{var}((X^\top X)^{-1}X^\top y|X) \\
&= [(X^\top X)^{-1}X^\top]\,\text{var}(y|X)[X(X^\top X)^{-1}] \\
&= [(X^\top X)^{-1}X^\top]\,[\sigma^2 I]\,[X(X^\top X)^{-1}] \\
&= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\
&= \sigma^2(X^\top X)^{-1}.
\end{aligned}$$

Ta thấy phương sai của $\hat{\theta}$ chỉ phụ thuộc vào X mà không phụ thuộc vào y.

Temperature and Pressure

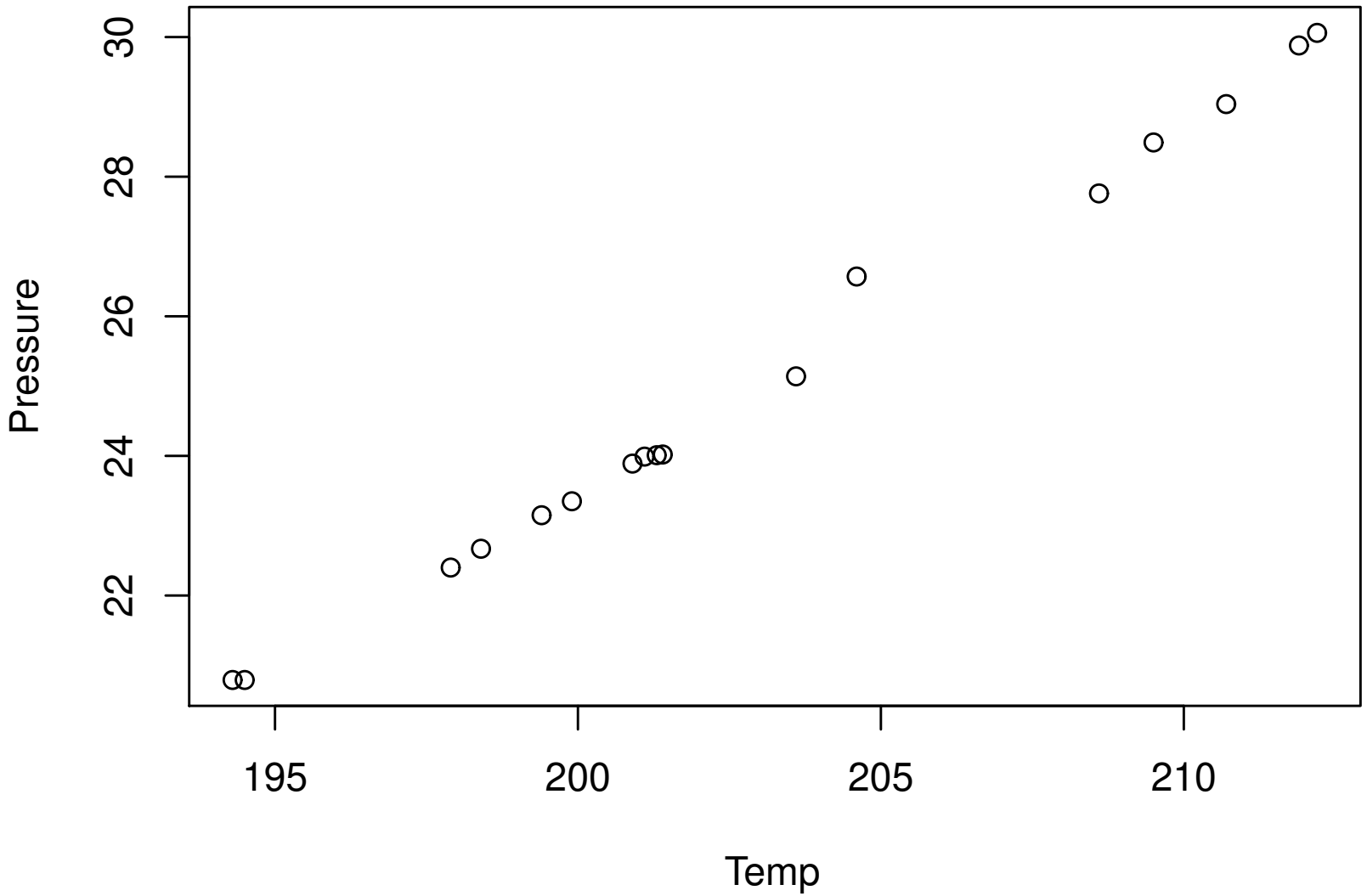
- Vào năm 1857, nhà vật lí học người Scotland James D. Forbes thực hiện một số thí nghiệm nhằm tìm hiểu mối liên hệ giữa áp suất và nhiệt độ sôi của nước.
- Ông biết rằng có thể xác định độ cao từ áp suất không khí đo bằng áp kế: càng lên cao áp suất càng thấp.
- Vào thời gian đó, áp kế thường có độ chính xác không cao. Forbes đề xuất thay thế áp kế bằng nhiệt độ sôi của nước.
- Forbes thu thập dữ liệu ở các dãy Alps và ở Scotland. Tại mỗi điểm, ông đo áp suất (theo inch thủy ngân) bằng áp kế và đo độ sôi của nước (theo độ Fahrenheit) bằng nhiệt kế.

Temperature and Pressure

Dữ liệu đo ở 17 điểm đo được cho trong bảng sau:

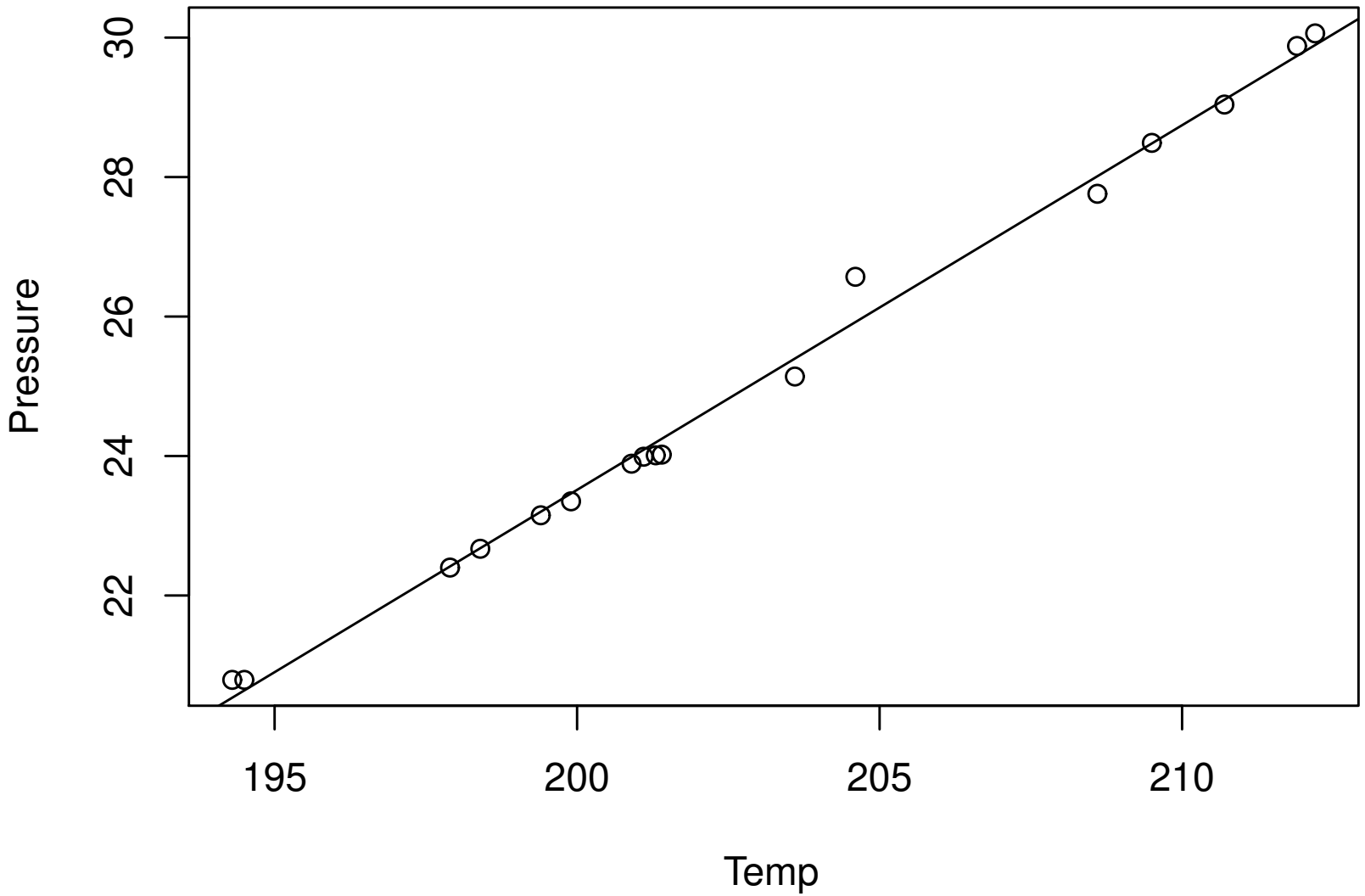
	Nhiệt độ	Áp suất
1	194.5	20.79
2	194.3	20.79
3	197.9	22.40
4	198.4	22.67
5	199.4	23.15
6	199.9	23.35
7	200.9	23.89
8	201.1	23.99
9	201.4	24.02
10	201.3	24.01
11	203.6	25.14
12	204.6	26.57
13	209.5	28.49
14	208.6	27.76
15	210.7	29.04
16	211.9	29.88
17	212.2	30.06

Temperature and Pressure



Temperature and Pressure

Ước lượng hồi quy cho kết quả: $h_\theta(x) = -81.06373 + 0.52289x$



Fuel Consumption

Trong ví dụ này, ta sử dụng mô hình hồi quy tuyến tính để

- Dự báo mức độ tiêu thụ nhiên liệu trong 50 bang của Hoa Kỳ và quận Columbia.
- Tìm hiểu hiệu ứng của tiêu thụ nhiên liệu đối với thuế xăng của các bang.

Fuel Consumption

Các biến dự báo được sử dụng trong ví dụ[1]. Dữ liệu được thu thập bởi Cục Đường bộ Hoa Kỳ vào năm 2001.

`Drivers`	Số bằng lái được cấp phép trong bang
`FuelC`	Lượng xăng sử dụng cho giao thông đường bộ, theo ngàn gallons
`Income`	Thu nhập bình quân đầu người năm 2000, theo ngàn đôla
`Miles`	Số dặm đường cao tốc của bang được hỗ trợ từ liên bang
`Pop`	Dân số lớn hơn hoặc bằng 16 tuổi
`Tax`	Thuế xăng của bang, theo cents trên một gallon
`State`	Tên bang
`Fuel`	$1000 \times$ `FuelC`/`Pop`
`Dlic`	$1000 \times$ `Drivers`/`Pop`
log(`Miles`)	Loga cơ số 2 của `Miles`

[1] Applied Linear Regression, 3rd edition, Sanford Weisberg, Wiley-Interscience, 2005.

Fuel Consumption

Một số thống kê mô tả dữ liệu tiêu thụ nhiên liệu:

Biến	N	Trung bình	Độ lệch chuẩn	Nhỏ nhất	Trung vị	Lớn nhất
`Tax`	51	20.15	4.5447	7.5	20.0	29.0
`Dlic`	51	903.7	72.858	700.2	909.1	1075.3
`Income`	51	28404	4451.637	20993	27871	40640
`logMiles`	51	15.75	1.4867	10.58	16.27	18.20
`Fuel`	51	613.1	88.96	317.5	626.0	842.8

$$\mathbb{E}(\texttt{Fuel}|X) = \theta_0 + \theta_1\texttt{Tax} + \theta_2\texttt{Dlic} + \theta_3\texttt{Income} + \theta_4\texttt{logMiles}.$$

Fuel Consumption

Ma trận thiết kế X và véc-tơ y là

$$X = \begin{pmatrix} 1 & 18.00 & 1031.38 & 23471 & 16.5271 \\ 1 & 8.00 & 1031.641 & 30064 & 13.7343 \\ 1 & 18.00 & 908.597 & 25578 & 15.7536 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 25.65 & 904.8936 & 21915 & 15.1751 \\ 1 & 27.30 & 882.329 & 28232 & 16.7817 \\ 1 & 14.00 & 970.7526 & 27230 & 14.7362 \end{pmatrix} \quad y = \begin{pmatrix} 690.264 \\ 514.279 \\ 621.475 \\ \vdots \\ 562.411 \\ 571.794 \\ 842.792 \end{pmatrix}$$

Các cột của ma trận thiết kế tương ứng với hệ số chắn, `Tax`, `Dlic`, `Income` và `logMiles`. Ma trận X có cỡ 51×5, còn y có cỡ 51×1.

Fuel Consumption

Ước lượng hồi quy cho kết quả như sau:

Biến	Hệ số	Sai số chuẩn
(Intercept)	154.192845	194.906161
Tax	−4.227983	2.030121
Dlic	0.471871	0.128513
Income	−0.006135	0.002194
logMiles	18.545275	6.472174

Linear Regression for Classification

- Mô hình hồi quy tuyến tính được sử dụng để phân loại theo quy tắc:
  - Sau khi ước lượng được tham số θ của mô hình hồi quy, với mỗi x ta tính

$$\hat{y} = h_\theta(x) = \sum_{j=1}^{D} \theta_j x_j.$$

  - Sau đó tùy thuộc vào giá trị của $\hat{y}$ mà ta quyết định phân x vào lớp nào.
- Với bài toán phân loại nhị phân, vì $y \in \{0, 1\}$ nên ta có thể sử dụng quy tắc phân loại sau: xếp x vào lớp 0 nếu $\hat{y} < 0.5$ và xếp x vào lớp 1 nếu $\hat{y} \geq 0.5$.

Breast Cancer

- Mỗi mẫu cần chẩn đoán có 30 đặc trưng:
  - Đặc trưng thứ nhất là trung bình của radius
  - Đặc trưng thứ 11 là độ lệch chuẩn của radius
  - Đặc trưng thứ 21 là radius lớn nhất, được tính bởi giá trị trung bình của 3 giá trị lớn nhất.
- Hai mẫu ví dụ được gán các lớp tương ứng là *ác tính* (*M*–malignant) và *lành tính* (*B*–benign).
  1. M, 17.99, 10.38, 122.8, 1001, 0.1184, 0.2776, 0.3001, 0.1471, 0.2419, 0.07871, 1.095, 0.9053, 8.589, 153.4, 0.006399, 0.04904, 0.05373, 0.01587, 0.03003, 0.006193, 25.38, 17.33, 184.6, 2019, 0.1622, 0.6656, 0.7119, 0.2654, 0.4601, 0.1189
  2. B, 7.76, 24.54, 47.92, 181, 0.05263, 0.04362, 0, 0, 0.1587, 0.05884, 0.3857, 1.428, 2.548, 19.15, 0.007189, 0.00466, 0, 0, 0.02676, 0.002783, 9.456, 30.37, 59.16, 268.6, 0.08996, 0.06444, 0, 0, 0.2871, 0.07039

Breast Cancer

Mô hình hồi quy tuyến tính sử dụng 10 đặc trưng:

$$
\begin{array}{rrrr}
\theta_0 = & 3.0521 & & \\
\theta_1 = & -0.49 & \theta_2 = & -0.022 \\
\theta_3 = & 0.055 & \theta_4 = & 0.001 \\
\theta_5 = & -1.9409 & \theta_6 = & -0.0973 \\
\theta_7 = & -0.8098 & \theta_8 = & -6.431 \\
\theta_9 = & -1.0119 & \theta_{10} = & 0.1193
\end{array}
$$

- Mô hình cho độ chính xác trên tập dữ liệu là 93.14%.
- Nếu sử dụng toàn bộ 30 đặc trưng thì độ chính xác của mô hình là 96.48%.