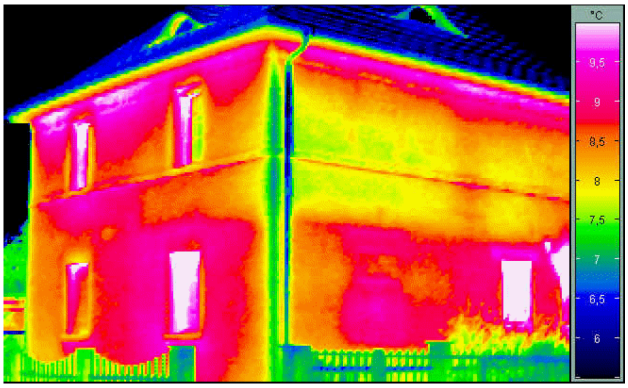
Example

Example: how does gas consumption depend on external temperature? (Whiteside, 1960s).



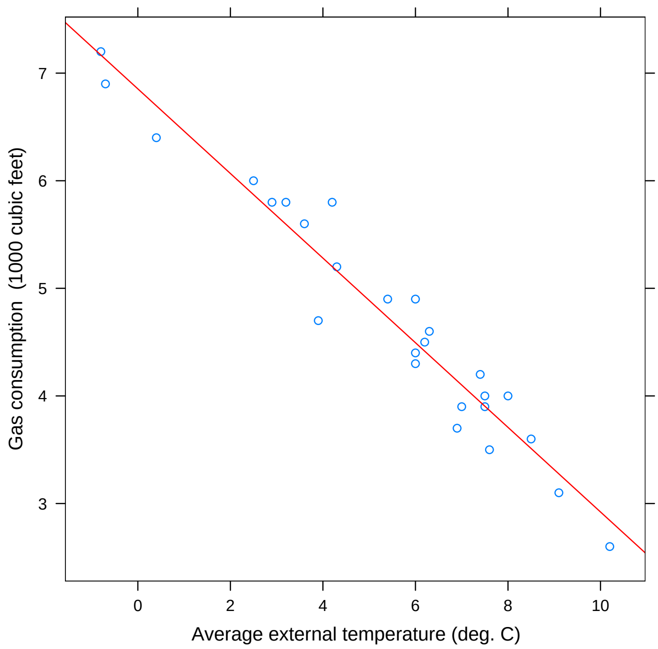
weekly measurements of

- average external temperature
- total gas consumption (in 1000 cubic feets)

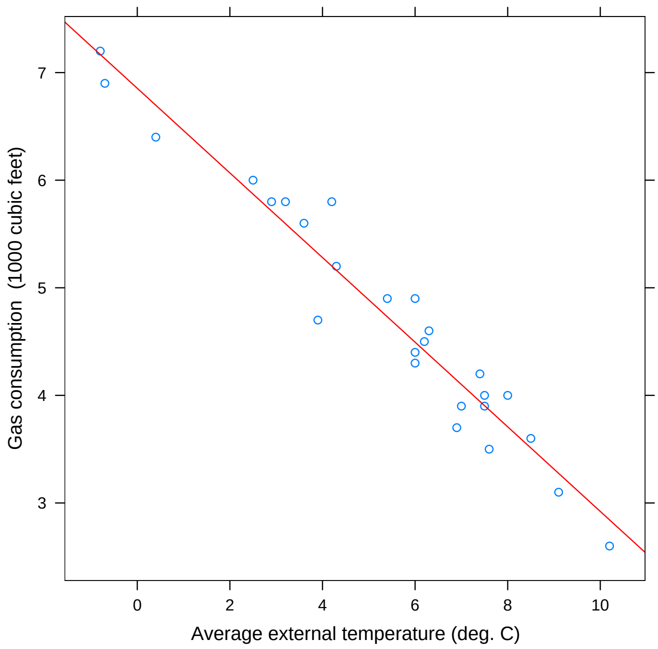
A third variable encodes two heating seasons, before and after wall insulation.

How does gas consumption depend on external temperature?

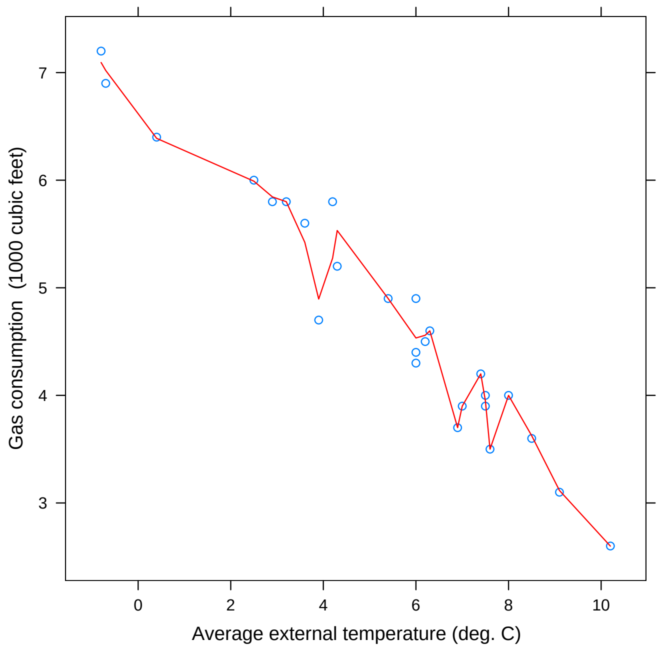
How much gas is needed for a given termperature ?



linear model



linear model



more flexible model

Variable Types and Coding

The most common variable types:

numerical / interval-scaled / quantitative
where differences and quotients etc. are meaningful,
usually with domain $\mathcal{X} := \mathbb{R}$,
e.g., temperature, size, weight.

nominal / discret / categorical / qualitative / factor
where differences and quotients are not defined,
usually with a finite, enumerated domain,
e.g., $\mathcal{X} := \{\text{red}, \text{green}, \text{blue}\}$
or $\mathcal{X} := \{\text{a}, \text{b}, \text{c}, \ldots, \text{y}, \text{z}\}$.

ordinal / ordered categorical
where levels are ordered, but differences and quotients are not defined,
usually with a finite, enumerated domain,
e.g., $\mathcal{X} := \{\text{small}, \text{medium}, \text{large}\}$

Variable Types and Coding

Nominals are usually encoded as binary **dummy variables**:

$$\delta_{x_0}(X) := \begin{cases} 1, & \text{if } X = x_0, \\ 0, & \text{else} \end{cases}$$

one for each $x_0 \in X$ (but one).

Example: $\mathcal{X} := \{\text{red}, \text{green}, \text{blue}\}$

Replace

one variable X with 3 levels: $\text{red}, \text{green}, \text{blue}$

by

two variables $\delta_{\text{red}}(X)$ and $\delta_{\text{green}}(X)$ with 2 levels each: 0, 1

X	$\delta_{\text{red}}(X)$	$\delta_{\text{green}}(X)$
red	1	0
green	0	1
blue	0	0
—	1	1

The Regression Problem Formally

Let

$X_1, X_2, \ldots, X_p$ be random variables called **predictors** (or **inputs**, **covariates**).
Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ be their domains.
We write shortly

$$X := (X_1, X_2, \ldots, X_p)$$

for the vector of random predictor variables and

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times X_p$$

for its domain.

Y be a random variable called **target** (or **output**, **response**).
Let $\mathcal{Y}$ be its domain.

$\mathcal{D} \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a (multi)set of instances of the unknown joint distribution $p(X, Y)$ of predictors and target called **data**.
$\mathcal{D}$ is often written as enumeration

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

The Regression Problem Formally

The task of regression and classification is
to predict Y based on X,
i.e., to estimate

$$r(x) := E(Y \mid X = x) = \int y \, p(y|x) dx$$

based on data (called **regression function**).

If Y is numerical, the task is called **regression**.

If Y is nominal, the task is called **classification**.