Statistics

Week 11 Recitation, Logistic Regression

ESD, SUTD

Term 5, 2017

Linear Regression Revisited

Probabilistic set up:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \text{Assume} \quad \epsilon_i \sim N(0, \sigma^2)$$

To estimate the coefficients β_0 and β_1:

Method I - Least Squares
to minimize the sum of squared errors:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Method II - Maximum Likelihood
to miximize the joint probability density:

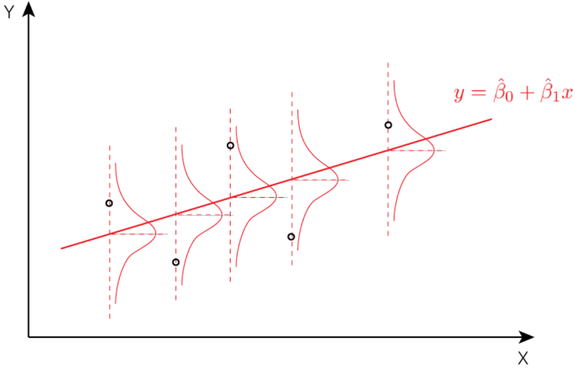
$$L = \prod_{i=1}^{n} \big(P(y_i|\beta_0, \beta_1)\big) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{\big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{2\sigma^2} \right) \right)$$

These two methods produce equivalent estimators.

Linear Regression Revisited

Reminder: the PDF of a normal distribution is

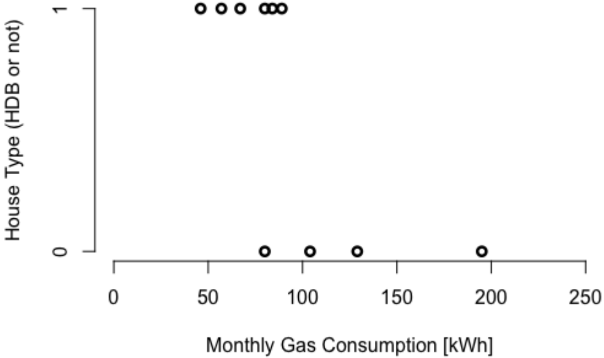
$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$$



Logistic Regression - Binary Dependent Variable

What if the dependent variable Y is binary?

Example: Using monthly gas consumption to decide if a residence is a HDB housing or not. 1-Yes, 0-No.

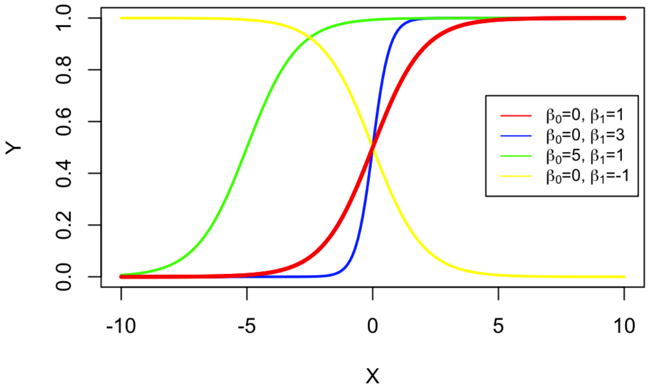


Logistic Regression - Modeling Probability

One solution:

To maximize the joint **Probability** of observing the sample data, where the probability of each observation being 1 is

$$P(y_i = 1|\beta_0, \beta_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$



Logistic Regression - Logistic Function

Standard logistic function:

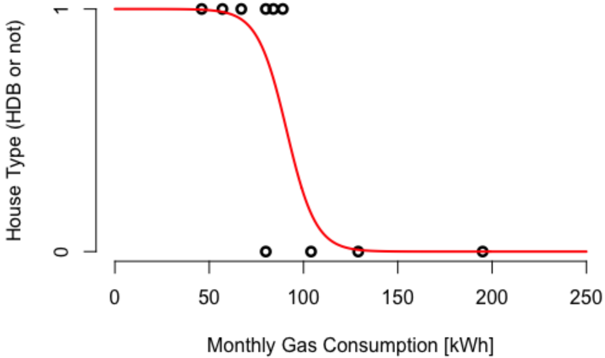
$$f(x) = \frac{1}{1 + e^{-x}}$$

Favorable properties:

- Bounded by [0,1], consistent with the concept of **probability**
- “S” curve
- $1 - f(x) = f(-x)$
- Derivative: $\frac{d}{dx} f(x) = f(x)\big(1 - f(x)\big)$

Logistic Regression - Coefficient Estimation

Back to the type of housing example:



Coefficient estimation: maximize the joint probability (likelihood):

$$L = \prod_{i=1}^{n} \left(P(y_i | \beta_0, \beta_1) \right) = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1 - y_i}$$

Logistic Regression - As a Generalized Linear Model

Probability function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

After transformation:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Logistic regression is a **Generalized Linear Model**.

$\dfrac{p}{1-p}$ is called the *odds ratio*:

It is the ratio of $P(y = 1)$ against $P(y = 0)$

Logistic Regression - Multiple Independent Variables

What if we have more than one independent variable, for example, three independent variables x_1, x_2 and x_3?

Solution:

$$P(y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}} = \frac{1}{1 + e^{-\mathbf{X_i}\beta}}$$

What if an independent variable is categorical, instead of numerical?

Solution: Dummy variables.

Logistic Regression - Exercise with R

Refer to the file ‘credit.csv’, which records people’s creditability and a set of other attributes (see text file ‘credit_description’ for details). To minimize the risk and maximize the profit of the bank, you are asked to fit a model to use the other attributes to predict the creditability.

Data source: Lichman, M. (2013). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)]

Useful Functions

```
factor(x)   # convert categorical variables to Factors
lr <- glm(formula, family=binomial, data)
# generalized linear model
step(lr)   # stepwise model selection by AIC
```

Note: R doesn’t automatically consider the interaction terms between dummy variables and continuous variables. You need to specify yourself.

Logistic Regression - Extension

What if the dependent variable y is a proportion?

Hint: the probability follows a binomial distribution

What if the dependent variable y has more than two categories?

Hint: multinomial logistic regression.

Logistic Regression - Summary

- Logistic regression is a regression model when the dependent variable y is binary (or proportional, or categorical)
- Logistic regression uses a function bounded within $[0, 1]$ to model the *probability* of y
- Logistic regression is a generalized linear model
- Logistic regression estimates coefficients using the maximum likelihood method