

Chapter 12: Multiple Regression and the General Linear Model

February 1, 2009

1 The multiple regression model

Consider the data shown in Figure 1, which gives the yields (in bushels) for 14 equal-sized plots planted in tomatoes for different levels of fertilization. It is evident that a simple linear regression model is not appropriate for the data. A model for this physical situation might be

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

which satisfies the following assumption:

- The mathematical form of the relation is correct, so $E(\epsilon_i) = 0$ for all i .
- $\text{Var}(\epsilon_i) = \sigma^2$ for all i .
- The ϵ_i s are independent.

- ϵ_i is normally distributed.

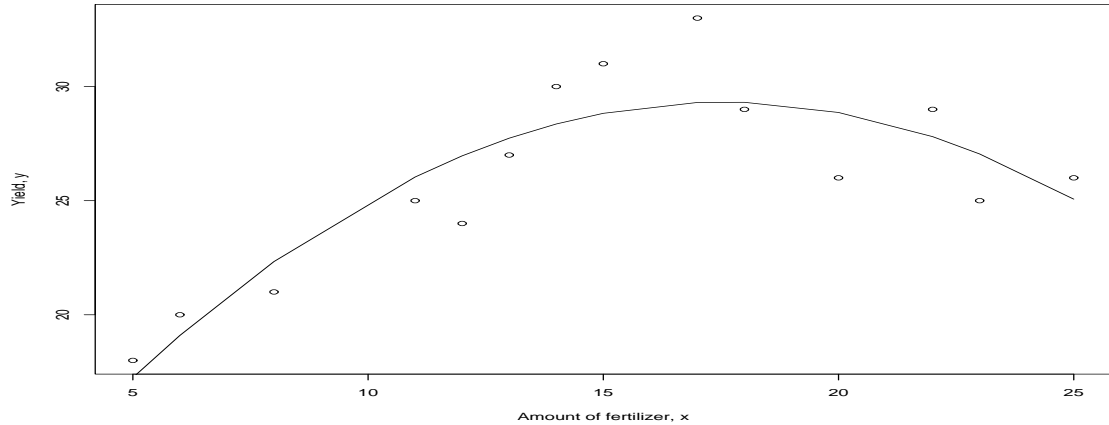


Figure 1: Scatterplot of the yield versus fertilizer data.

The multiple regression model, which relates a response variable y to a set of k quantitative explanatory variables, is a direct extension of the polynomial regression model in one independent variable. The multiple regression model is expressed as

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon.$$

Any of the k explanatory variables may be powers of the independent variables, for example, $x_3 = x_1^2$, or a cross-product term, $x_4 = x_1 x_2$, or a nonlinear function such as $x_5 = \log(x_1)$, and so on. So this model is also called the general linear model. The word “linear” refers to how the β s are entered in the model, not to how the independent

variables appear in the model. For example, the following two models

$$y = \beta_0 + \beta_1 x^2 + \epsilon$$

and

$$y = \beta_0 + \beta_1 \sin(x_1) + \beta_2 \log(x_2) + \epsilon$$

are still linear models. While the model

$$y = \beta_1 x_1 \exp(\beta_2 x_2) + \epsilon$$

is non-linear.

Consider the problem of writing a model for an experimental situation in which a response y is related to a set of qualitative independent variables or to both quantitative and qualitative independent variables. For example, in an experiment we are interested in four levels of qualitative variables. We call these levels treatments. We would write the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

where

$$x_1 = 1 \text{ if treatment 2, } x_1 = 0 \text{ otherwise,}$$

$$x_2 = 1 \text{ if treatment 3, } x_2 = 0 \text{ otherwise,}$$

$$x_3 = 1 \text{ if treatment 4, } x_3 = 0 \text{ otherwise.}$$

Treatment			
1	2	3	4
$E(y) = \beta_0$	$E(y) = \beta_0 + \beta_1$	$E(y) = \beta_0 + \beta_2$	$E(y) = \beta_0 + \beta_3$

To interpret the β s in this equation, we construct a table of the expected values:

If we identify the mean of treatment 1 as μ_1 , the mean of treatment 2 as μ_2 , and so on. Then we have

$$\mu_1 = \beta_0, \mu_2 = \beta_0 + \beta_1, \mu_3 = \beta_0 + \beta_2, \mu_4 = \beta_0 + \beta_3.$$

2 Model Estimation

In matrix notation, the multiple linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We wish to find the vector of least squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

The least squares estimators must satisfy

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

which simplifies to

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

The above equations are the least squares normal equations. The least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provided that the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ exists. The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ will always exist if the regressors are linearly independent, that is, if no column of the \mathbf{X} matrix is a linear combination of the other columns.

The fitted regression model corresponding to the levels of the regressor variables $\mathbf{x}' = [1, x_1, x_2, \dots, x_k]$ is

$$\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j.$$

The vector of fitted values \hat{y}_i is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is usually called the **hat matrix**. The n residuals may be conveniently written as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

2.1 Properties of the least squared estimator

- If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.
- If $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the least squares estimators $\hat{\beta}_j$, $j = 0, 1, \dots, k$, have minimum variance among all linear unbiased estimators.

2.2 Estimation of σ^2

The residual sum of squares

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

The residual mean square is

$$MS_{Res} = \frac{SS_{Res}}{n - k - 1},$$

and it is an unbiased estimate of σ^2 .

2.3 Example

Fuel Consumption The goal of this example is to understand how fuel consumption varies over the 50 United States. The variables considered in this example are as follows.

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
State	State name

Fuel	$1000 \times \text{FuelC}/\text{Pop}$
Dlic	$1000 \times \text{Drivers}/\text{Pop}$
log(Miles)	Base-two logarithm of Miles

The scatterplot matrix for the fuel data is shown in Figure 2. Each

plot in a scatterplot matrix is relevant to a particular one-predictor regression of the variable on the vertical axis, given the variable on the horizontal axis.

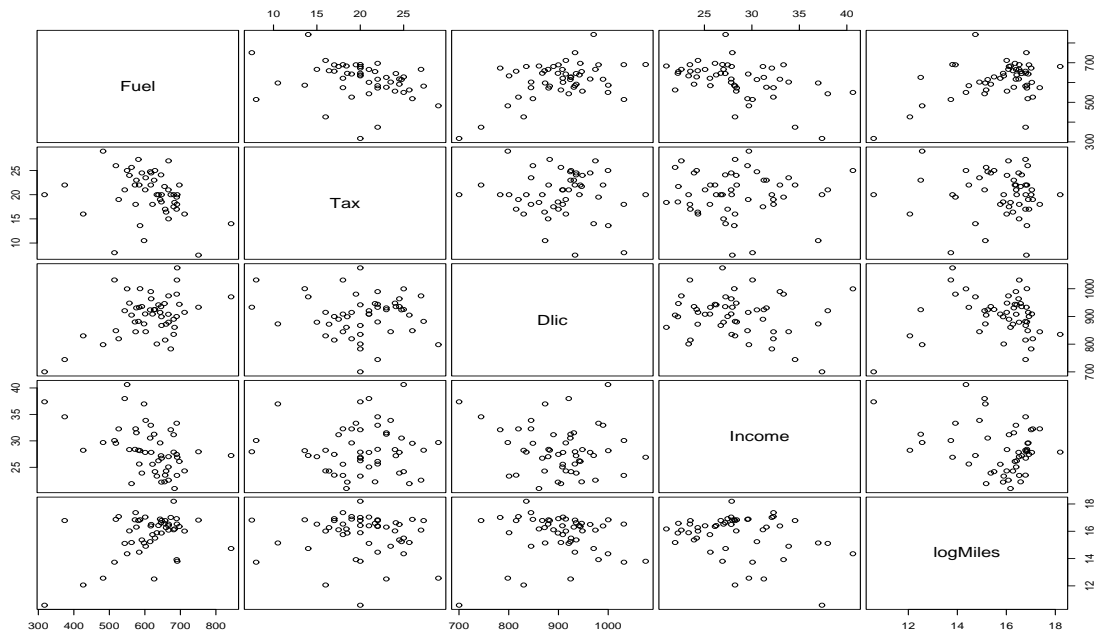


Figure 2: Scatterplot matrix for the fuel data.

A more traditional and less informative, summary of the two-variable relationships is the matrix of sample correlations, shown in Table 3.2. In this instance, the correlation matrix helps to reinforce the relationships we see in the scatterplot matrix, with fairly small correlations between the predictors and Fuel, and essentially no correlation between the predictors themselves.

Fit the fuel data by a multiple linear regression model with mean

Table 1: Sample correlation for the fuel data.

	Tax	Dlic	Income	logMiles	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

function

$$E(\text{Fuel}|X) = \beta_0 + \beta_1 \text{Tax} + \beta_2 \text{Dlic} + \beta_3 \text{Income} + \beta_4 \log(\text{Miles}).$$

The 5×5 matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is given by

	Intercept	Tax	Dlic	Income	logMiles
Intercept	9.02151	-2.852e-02	-4.080e-03	-5.981e-02	-1.933e-01
Tax	-0.02852	9.788e-04	5.599e-06	4.263e-05	1.602e-04
Dlic	-0.00408	5.599e-06	3.922e-06	1.189e-05	5.402e-06
Income	-0.05981	4.263e-05	1.189e-05	1.143e-03	1.000e-03
logMiles	-0.19315	1.602e-04	5.402e-06	1.000e-03	9.944e-02

The coefficients can then be calculated as

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (154.1928, -4.2280, 0.4719, -6.1353, 18.5453)'$$

The output gives the estimates $\hat{\beta}$ and their standard errors computed based on $\hat{\sigma}^2$ and the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	154.1928	194.9062	0.791	0.432938
Tax	-4.2280	2.0301	-2.083	0.042873
Dlic	0.4719	0.1285	3.672	0.000626
Income	-6.1353	2.1936	-2.797	0.007508
logMiles	18.5453	6.4722	2.865	0.006259

Residual standard error: 64.89 on 46 degrees of freedom

Multiple R-Squared: 0.5105, Adjusted R-squared: 0.4679

F-statistic: 11.99 on 4 and 46 DF, p-value: 9.33e-07

3 The analysis of variance

The test for significance of regression is a test to determine if there is a linear relationship between the response y and any of the regressor variables x_1, x_2, \dots, x_k . This procedure is often thought of as an overall or global test of model adequacy. The appropriate hypothesis are

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \quad \text{for at least one } j$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regress	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n - k - 1$	MS_{Res}	
Total	SS_T	$n - 1$		

Table 2: Analysis of Variance (ANOVA) for testing significance of regression

Rejection of this null hypothesis implies that at least one of the regressors x_1, \dots, x_k contributes significantly to the model.

The test is based on the identity:

$$SS_T = SS_R + SS_{Res},$$

where $SS_R = \hat{\beta}' \mathbf{X}' \mathbf{y} - n \bar{\mathbf{y}}^2$, $SS_{Res} = \mathbf{y}' \mathbf{y} - \hat{\beta}' \mathbf{X}' \mathbf{y}$, and $SS_T = \mathbf{y}' \mathbf{y} - n \bar{\mathbf{y}}^2$; and the following ANOVA table. Therefore, to test the null hypothesis, compute the test statistic F_0 and reject H_0 if $F_0 > F_{\alpha, k, n-1}$.

The overall ANOVA table for the fuel data is given by

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regress	201994	4	50499	11.992
Residual	193700	46	4211	
Total	395694	50		

To get a significance level for the test, we would compare $F_0 =$

11.992 with the $F(4, 46)$ distribution. Since the probability $Pr(> F_0) = 9.33e - 07$, a very small number, leading to a very strong evidence against the null hypothesis that the mean function does not depend on any of the terms. The value of $R^2 = 201994/395694 = 0.5105$ indicates that about half the variation in Fuel is explained by the terms.

3.1 R^2 and Adjusted R^2

$$R^2 = 1 - \frac{SS_{Res}}{SS_T}$$

$$R_{Adj}^2 = 1 - \frac{SS_{Res}/(n - p)}{SS_T/(n - 1)}$$

The adjusted R^2 penalizes us for adding terms that are not helpful, so it is very useful in evaluating and comparing candidate regression models.

3.2 Tests on individual regression coefficients

The hypotheses for testing the significance of any individual regression coefficient, such as β_j , are

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_j \neq 0$$

If H_0 is not rejected, then this indicates that the regressor x_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)},$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. The null hypothesis is rejected if $|t_0| > t_{\alpha/2, n-k-1}$. Note that this is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variables $x_i (i \neq j)$ that are in the model. Thus, this is a test of the contribution of x_j given the other regressors in the model.

Consider the regression model with k regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

where $p = k + 1$, $\boldsymbol{\beta}_1$ is a $(p - r)$ -vector of coefficients, and $\boldsymbol{\beta}_2$ is a r -vector of coefficients. We wish to test the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

To find the contribution of the terms in $\boldsymbol{\beta}_2$ to the regression, fit the model assuming that the null hypothesis H_0 is true. The reduced model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The LS estimator of $\boldsymbol{\beta}_1$ in the reduced model is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$.

The regression sum of squares is

$$SS_R(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1 \mathbf{X}'_1 \mathbf{y} - \left(\sum_{i=1}^n y_i \right)^2 / n$$

The regression sum of squares due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is

$$SS_R(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1)$$

with $p - (p - r) = r$ degrees of freedom. This sum of squares is called the **extra sum of squares due to $\boldsymbol{\beta}_2$** because it measures the increase in the regression sum of squares that results from adding the regressors \mathbf{X}_2 to a model that already contains \mathbf{X}_1 . Now $SS_R(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1)$ is independent of MS_{Res} , and the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1) / r}{MS_{Res}}$$

If $F_0 > F_{\alpha, r, n-p}$, we reject H_0 , concluding that at least one of the parameters in $\boldsymbol{\beta}_2$ is not zero, and consequently at least one of the regressors x_{k-r+1}, \dots, x_k in \mathbf{X}_2 contribute significantly to the regression model. This test is also a partial F test because it measures the contribution of the regressors in \mathbf{X}_2 given that the other regressors in \mathbf{X}_1 are in the model.

Analysis of Variance Table

Model 1: Fuel ~ Dlic + Income + logMiles

Model 2: Fuel ~ Tax + Dlic + Income + logMiles

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	211964				
2	46	193700	1	18264	4.3373	0.04287

Note that the t -statistic for Tax is $t = -2.083$, and $t^2 = (-2.083)^2 = 4.34$, the same as the F -statistic we just computed.

4 Confidence interval: estimation of the mean response

We may construct a confidence interval on the mean response at a particular point $\mathbf{x}_0 = (1, x_{01}, \dots, x_{0k})'$. The fitted value at this point is

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$$

This is an unbiased estimate of $E(y|\mathbf{x}_0)$, and the variance of \hat{y}_0 is

$$\text{Var}(\hat{y}_0) = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

Therefore, a $100(1 - \alpha)$ percent confidence interval on the mean response at the point \mathbf{x}_0 is

$$\hat{y}_0 - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0} \leq E(y|\mathbf{x}_0) \leq \hat{y}_0 + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0}$$

5 Prediction of new observations

A point estimate of the future observation y_0 at the point \mathbf{x}_0 is

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}.$$

A $100(1 - \alpha)$ percent prediction interval for this future observation is

$$\hat{y}_0 - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0)} \leq E(y|\mathbf{x}_0) \leq \hat{y}_0 + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0)}$$

6 Logistic Regression

When the response variable y is binary, the distribution of y reduces to a single value, the probability $p = Pr(y = 1)$. Let $p(x)$ be the probability that y equals 1 when the independent variable equals x .

We can model the log-odds ratio to a linear model in x , a simple

logistic regression model:

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

which transforms to

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

The parameters β_0 and β_1 of the model can be estimated using the maximum likelihood method; that is, maximizing the function

$$L(\beta_0, \beta_1 | y_1, \dots, y_n; x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1 - y_i}.$$