Regression

Sunita Sarawagi

CS 215. Fall 2024

Reading: Chapter 9 in Ross Textbook

Problem definition

- So far, we have been estimating the density of single random variables $f(X)$
- In many applications, we need to reason about the distribution of values of a continuous random variable Y but as a function of some other variables $\mathbf{x} = (x_1, \ldots, x_k)$
  - Y is called output or response or dependent variable
  - $(x_1, \ldots, x_k)$ are called input or covariate or independent variables.

- We want to estimate the conditional density: $f(Y|\, x_1, \ldots, x_k)$
- We are given data samples D of the form:
- $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} = \{(x_{i1}, x_{i2}, \ldots x_{ik}, y_i) : i = 1, \ldots, n\}$

Motivating examples

- How does errors in assembled circuits (Y) depend on temperature (x1), percentage of copper (x2), humidity (x3)

  Yield

  f (yield | temperature, % copper, humidity)

- Predict stock price tomorrow (Y) as a function of stock price in the last 7 days $(x_1, x_2, \dots, x_7)$

- Express CPU temperature (Y) as a function of workload (x1), ambient temperature (x2), fan-speed (x3), chip model (x4), etc.

How to represent conditional density?

- Many different forms of f(Y|**x**) will be discussed later.
- A simple form:

$$f(Y|\, x_1, \ldots, x_k) \sim N(\mu_x, \sigma^2), \quad \text{where } \mu_x = \beta_1 x_1 + \cdots + \beta_k x_k + \alpha$$

- Parameters of the above model are

$$\beta_1, \beta_2 \ldots \beta_k, \alpha, \sigma^2$$

- Alternate way to view the above:

$$Y = \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \alpha}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{random error}} \qquad \varepsilon \sim N(0, \sigma^2)$$

Special case k = 1

$$f(Y \mid x_1) \sim N(\beta_1 x_1 + \alpha \,;\, \sigma^2)$$

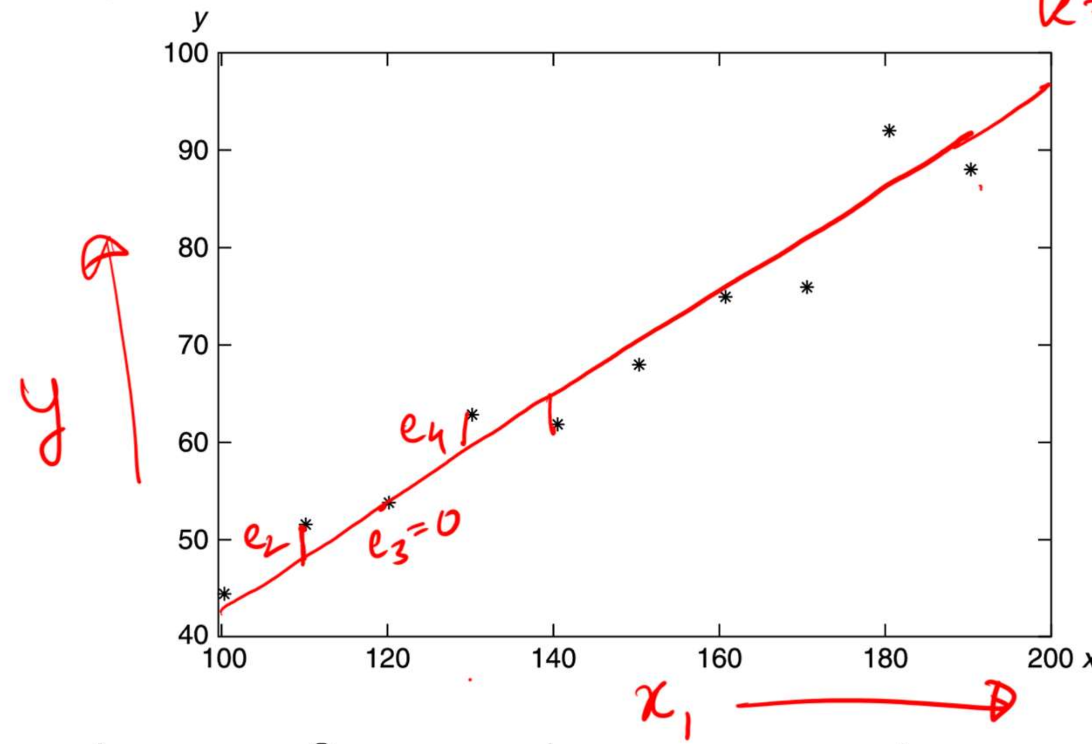
$$Y = \beta_1 x_1 + \alpha + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

Example 9.1.a. Consider the following 10 data pairs (x_i, y_i), $i = 1, \ldots, 10$, relating y, the percent yield of a laboratory experiment, to x, the temperature at which the experiment was run.

$n = 10$
$k = 1$

D

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88



Our goal is to use the data to learn α, β, σ so that we can write

$$y_i = \beta x_i + \alpha + e_i, \qquad e_i \sim N(0, \sigma^2)$$

Notation

Actual unknown parameter β, α, σ^2

Estimated parameters using n samples.

$\hat{\beta}_n$ $\hat{\alpha}_n$ $\hat{\sigma}_n$ } previous notation

Instead following the Ross textbook notation.

B, A, S

β α σ

Estimating parameters from data using Maximum Likelihood

- Given data D $= \{(x_1, y_1), \ldots (x_n, y_n)\} = \{(x_i, y_i) ; i = 1 \ldots n\}$
- Theorem: The parameters β, α that maximize the likelihood of D are:

$$B = \frac{\sum_i x_i Y_i - n\overline{x}\overline{Y}}{\sum_i x_i^2 - n\overline{x}^2} \qquad A = \overline{Y} - B\overline{x}$$

[Ignore estimation of σ^2]

Proof:

Log likelihood of the data:

$$\max_{A,B} LL(D) = \sum_{i=1}^{n} \log P(y_i | x_i) = \sum_{i=1}^{n} \log \frac{e^{-\frac{(\overbrace{y_i - Bx_i}^{u_x} - A)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

$$\equiv \min_{A,B} \sum_{i=1}^{n} (y_i - Bx_i - A)^2$$

$$\frac{\partial LL}{\partial A} = \sum_{i=1}^{n} (y_i - Bx_i - A)(-1) = 0$$

$$\Rightarrow nA = \sum_{i=1}^{n} y_i - B\sum_{i=1}^{n} x_i$$

$$\Rightarrow A = \bar{y} - B\bar{x} \quad \text{where } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

$$\frac{\partial LL}{\partial B} = \sum_{i=1}^{n} (y_i - Bx_i - A)(-x_i) = 0$$

$$\sum_{i=1}^{m} (y_i - Bx_i - \bar{y} + B\bar{x}) x_i = 0$$

$$\Rightarrow B\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_i x_i\right) = \sum x_i y_i - \sum_i x_i \bar{y}$$

$$B = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Analyzing risk of the linear regression parameters

We will show the following results about the mean and variance of the maximum likelihood estimators A, B of α, β respectively.

- $E[B] = \beta, \quad E[A] = \alpha$: Both are unbiased estimators.
- $Var[B] = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$ $\qquad \mathrm{Var}(A) = \dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$

More generally we can show that both parameters follow a Gaussian distribution with above mean and variance.

$$B \sim N\left(\beta, \frac{\sigma^2}{\sum_i x_i^2 - n\,\bar{x}^2}\right), \qquad A = N(\alpha, Var(A))$$

$$B = \frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$B = \sum_{i=1}^{n} Y_i \left[\frac{x_i - \bar{x}}{SS} \right] \qquad Y_i \sim N(\beta x_i + \alpha ; \sigma^2)$$

$$\Rightarrow B = \sum_{i=1}^{n} w_i Y_i \Rightarrow B \sim N\left(\sum_{i=1}^{n} w_i E[Y_i] ; \sum_{i=1}^{n} w_i^2 \operatorname{var}(Y_i) \right)$$

B is normally distributed.

$$E[B] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{SS} (\beta x_i + \alpha) = \beta$$

simple manipulations.

$$\frac{\sum_i (x_i - \bar{y})(\beta x_i + \alpha)}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\beta\left(\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)\right) + \alpha\left(\sum_i (x_i - \bar{x})\right)}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\alpha \sum_i x_i - \alpha n \bar{x} =$$

$$= \alpha n \bar{x} - \alpha n \bar{x} = 0$$

$$\frac{\beta\left[\sum_{i=1}^{n} x_i^2 - \bar{x} n \bar{x}\right]}{\sum_i x_i^2 - n\bar{x}^2} = \beta .$$

$$\mathrm{Var}[B] = \sum_i \mathrm{Var}(Y_i)\left[\frac{(x_i - \bar{x})^2}{\left(\sum_i x_i^2 - n\bar{x}^2\right)^2}\right] = \sigma^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\sum_i x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_i x_i = \sum_i x_i^2 - n\bar{x}^2$$

$$\mathrm{Var}(B) = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}$$

$$B \sim N\left(\beta \,;\, \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}\right) = N\left(\beta ; \frac{\sigma^2}{n\sigma_x^2}\right) \text{ where}$$

$$\sigma_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$

Distribution of A estimate:

$$A = \bar{Y} - B\bar{x}$$

$$E[A] = E[\bar{Y}] - E[B\bar{x}]$$

$$\mathrm{Var}[A] = \mathrm{Var}(\bar{Y}) + \mathrm{Var}(B\bar{x})$$ ← erroneous because $\bar{Y}$ is not independent of B..

$$A = \sum_{i=1}^{n} Y_i \underbrace{\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right)}_{\text{constants.}} \qquad A = \bar{Y} - B\bar{x}$$

$$A = \sum_{i=1}^{n} Y_i \sigma_i \implies A \sim N\left(\sum_i \sigma_i (\beta x_i + \alpha) \, ; \, \sigma^2 \sum_i \sigma_i^2 \right)$$

$$E[A] = E[\bar{Y}] - E[B\bar{x}] \, : \, \frac{1}{n} E\left(\sum_{i=1}^{n} Y_i \right) = \sum_{i=1}^{n} E[Y_i]/n$$

$$= [\beta \bar{x} + \alpha] - \bar{x}\beta = \alpha \qquad = \frac{\sum_{i=1}^{n} \beta x_i + \alpha}{n}$$

$$= \beta \bar{x} + \alpha$$

$\mathrm{Var}[A] =$ homework.

Estimating σ^2

- Calculate sum of square of residuals
  - Residuals = difference between actual y_i and predicted value $\mathrm{B}x_i + A$

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- The MLE estimate would be:

- The above biased like for normal Gaussian parameters.
- We will use a different method:

The Chi-Square distribution (Section 5.8 of textbook)

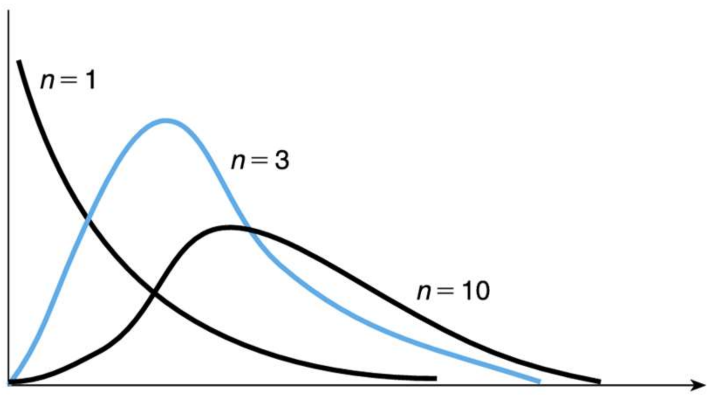
Definition. If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then X, defined by

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{5.8.1}$$

is said to have a *chi-square distribution with n degrees of freedom*. We will use the notation

$$X \sim \chi_n^2$$

to signify that X has a chi-square distribution with n degrees of freedom.



Deriving the density of χ_n^2 distribution

- Use MGF.
- Consider n=1 first.

$$
\begin{aligned}
E[e^{t\,X}] &= E[e^{t\,Z^2}] \text{ where } Z \sim \mathcal{N}(0,1) \\
&= \int_{-\infty}^{\infty} e^{t\,x^2} f_Z(x)\,dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t\,x^2} e^{-x^2/2}\,dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2(1-2t)/2}\,dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\bar{\sigma}^2}\,dx \quad \text{where } \bar{\sigma}^2 = (1-2t)^{-1} \\
&= (1-2t)^{-1/2} \frac{1}{\sqrt{2\pi}\,\bar{\sigma}} \int_{-\infty}^{\infty} e^{-x^2/2\bar{\sigma}^2}\,dx \\
&= (1-2t)^{-1/2}
\end{aligned}
$$

General n

- $E_X\left[e^{\{tX\}}\right] = E\left[e^{t\sum_i Z_i^2}\right] = \prod_i E\left[e^{tZ_i^2}\right] = (1-2t)^{-n/2}$

- The above is MGF of gamma distribution with parameters (n/2,1/2).

A random variable is said to have a gamma distribution with parameters (α, λ), $\lambda > 0$, $\alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Thus, density of χ^2 distribution is

- $$f(x) = \frac{\frac{1}{2}e^{-x/2}\left(\frac{x}{2}\right)^{(n/2)-1}}{\Gamma\left(\frac{n}{2}\right)}, \quad x > 0$$

Expected value of χ_n^2 distribution

- E[χ_n^2] = n   [Can be derived from the MGF]

- Var[χ_n^2] = 2n

Estimating σ^2

- Calculate sum of square of residuals where
  - Residuals = difference between actual y_i and predicted value $\mathrm{B}x_i + A$

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- It can be show that $\frac{SS_R}{\sigma^2}$ follows a Chi-square distribution with n-2 degrees of freedom
  - Book has a kind of intuitive proof…

Estimating σ^2

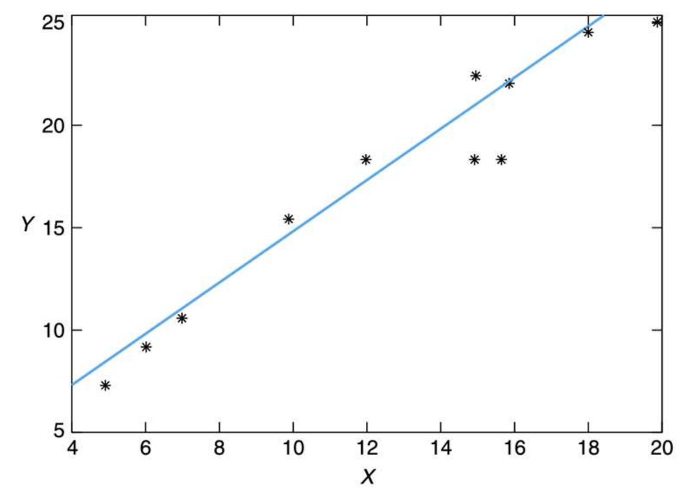
Let estimate of σ^2 be called S.

$$S = \frac{SS_R}{n-2}$$

S is an unbiased estimate of σ^2. It is easy to see that $E[S] = \sigma^2$

Example 9.3.a. The following data relate x, the moisture of a wet mix of a certain product, to Y, the density of the finished product.

x_i	y_i
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
20	24.8



Compute the least square fit. Estimate A, B, S

$$y = 2.463 + 1.206x$$

Multi-variable linear regression

Reading material: Section 9.10 of Ross Textbook

General case: k > 1

$$f(Y|\, x_1, \dots, x_k) \sim \;\; N(\mu_x, \sigma^2), \quad where \;\; \mu_x = \beta_1 x_1 + \cdots + \beta_r x_r + \beta_0$$

Or

Y = $\beta_1 x_1 + \cdots + \beta_r x_r + \beta_0$+ e   where $e \sim N(0, \sigma^2)$

Training data D will be denoted as

$$\{(x_{i1}, x_{i2}, \dots x_{ir}, y_i): i = 1 \dots n\}$$

MLE estimates of parameters:

Solving the MLE

Solving the MLE

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik} \tag{9.10.1}$$

$$\sum_{i=1}^{n} x_{i1} Y_i = B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{k} x_{ik} Y_i = B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}^2$$

Matrix notation for k-dimensional covariates.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

then $\mathbf{Y}$ is an $n \times 1$, $\mathbf{X}$ an $n \times p$, $\boldsymbol{\beta}$ a $p \times 1$, and $\mathbf{e}$ an $n \times 1$ matrix where $p \equiv k + 1$.

The regression problem now becomes: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_{i1}Y_i \\ \vdots \\ \sum_i x_{ik}Y_i \end{bmatrix}$$

Solving the MLE

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik} \tag{9.10.1}$$

$$\sum_{i=1}^{n} x_{i1} Y_i = B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{k} x_{ik} Y_i = B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}^2$$

$$\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}'$ is the transpose of $\mathbf{X}$.

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Example

Get least square estimate on this data with last column as y.

Table 9.5

	Age (years)	Elevation (1000 ft)	Rain-fall(inches)	Specific Gravity	Diameter at Breast Height (inches)
1	44	1.3	250	.63	18.1
2	33	2.2	115	.59	19.6
3	33	2.2	75	.56	16.6
4	32	2.6	85	.55	16.4
5	34	2.0	100	.54	16.9
6	31	1.8	75	.59	17.0
7	33	2.2	85	.56	20.0
8	30	3.6	75	.46	16.6
9	34	1.6	225	.63	16.2
10	34	1.5	250	.60	18.5
11	33	2.2	255	.63	18.7
12	36	1.7	175	.58	19.4
13	33	2.2	75	.55	17.6
14	34	1.3	85	.57	18.3
15	37	2.6	90	.62	18.8

$$y = 11.54873 + 0.05728\,x_1 + 0.08712\,x_2 + 7.33231\,x_3$$

Non-parametric regression