Linear Regression

A Two Variable Model

Noara Razzak

May 9, 2020

Linear Regression

The regression model is a statistical procedure that allows one to estimate the linear, or straight line, relationship that relates two or more variables. This linear relationship summarizes the amount of change in one variable that is associated with change in another variable or variables

Linear Regression

Key Points

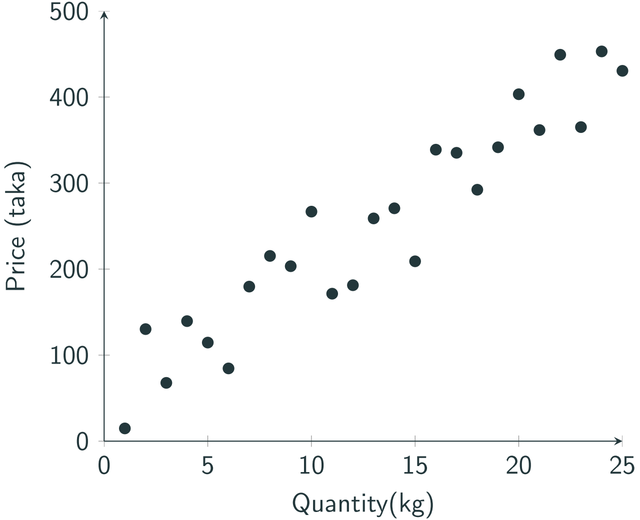
- The two variable regression model assigns one of the variables the status of an independent variable, and the other variable the status of a dependent variable.
- The independent variable may be regarded as causing changes in the dependent variable.
- However, one cannot be certain of a causal relationship, even with the regression model.

Linear Regression

Suppose we wanted to see how the price of rice per kilogram varies due to the quantity of rice bought.

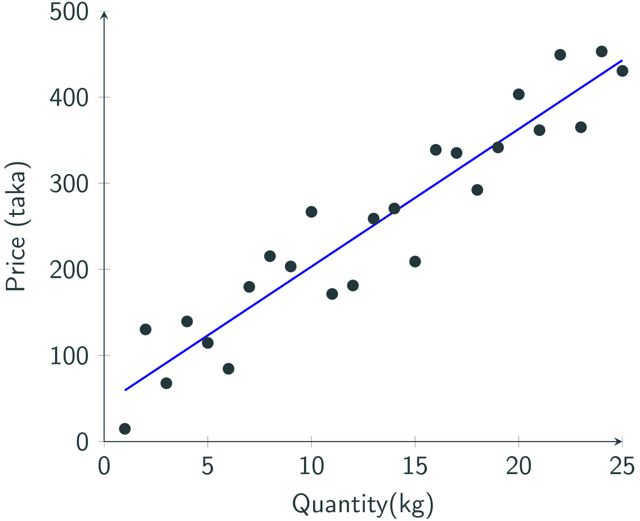
Linear Regression

An illustration



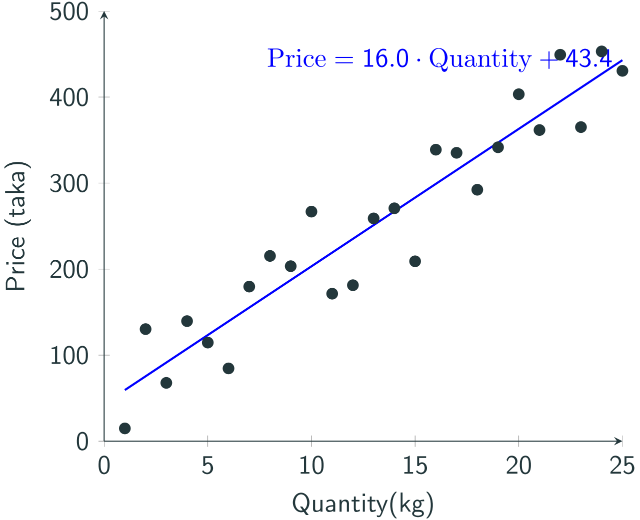
Linear Regression

An illustration



Linear Regression

An illustration



Linear Regression

Basic Model

We are going to fit a line $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$ to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable.

In our example, the variable Weight is the independent variable and the variable Price is the dependent variable.

Linear Regression

Explanation

- β_1 is the slope of the line: this is one of the most important quantities in any linear regression analysis. A value very close to 0 indicates little to no relationship; large positive or negative values indicate large positive or negative relationships.
- β_0 is the intercept of the line.

Linear Regression

Explanation

We observe paired data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where we assume that as a function of x_i, each y_i is generated by using some true underlying line $y_i = \beta_0 + \beta_1 \cdot x + \epsilon_i$ that we evaluate at x_i.

But notice that we are only fitting the best line we can through the points. Therefore, each point for which the true line is fitted will have an associated error.

Linear Regression

What is the error?

We will assume nothing concrete about the error except the fact that the error is normally distributed. Formally,we express the true model the following way:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here, the noise ϵ_i represents the fact that our data does not fit the model perfectly. Since ϵ_i is normally distributed we write that $\epsilon_i \sim N(0, \sigma^2)$.

Linear Regression

How do we solve the model?

Our aim is to find β_1, as it will help us determine how much we will need to spend to obtain a certain amount of rice. However, we will also need to calculate β_0 in the process.

Linear Regression

Optimization

- This is a simple optimization problem. Our aim is to minimize the error or ϵ_i as much as possible. Therefore, we subtract $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$ from $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- Basically we minimize the sum of all the $\epsilon_i^2 s$.
- Take a moment to think why we are minimizing the squared values.

Linear Regression

Optimization

We need to minimize the error squared to obtain as accurate a value of β_1 and β_0 as possible. Therefore we minimize the following function:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

$$= \min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 \cdot x)]^2$$

Linear Regression

Solving the optimization problem

Let's name the function $\sum_{i=1}^{n} \epsilon_i^2$ as the function Q.

Now the function Q will be minimized if the First Order Conditions are applied: $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$.

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_0:

$$
\begin{aligned}
&\frac{\partial Q}{\partial \beta_0} = 0 \\
&\Rightarrow \sum_{i=0}^{n} 2 \cdot (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-1) = 0 \\
&\Rightarrow -2 \cdot \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i) = 0 \\
&\Rightarrow -\sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i) = 0 \\
&\Rightarrow -\sum_{i=0}^{n} y_i + \beta_0 \sum_{i=0}^{n} \cdot 1 + \beta_1 \cdot \sum_{i=0}^{n} x_i = 0
\end{aligned}
$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_0:

$$\Rightarrow \beta_0 \sum_{i=0}^{n} \cdot 1 = \sum_{i=0}^{n} y_i - \beta_1 \cdot \sum_{i=0}^{n} x_i$$

$$\Rightarrow n \cdot \beta_0 = \sum_{i=0}^{n} y_i - \beta_1 \cdot \sum_{i=0}^{n} x_i$$

$$\Rightarrow \beta_0 = \frac{1}{n} \cdot \left[\sum_{i=0}^{n} y_i - \beta_1 \cdot \sum_{i=0}^{n} x_i \right]$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_1:

$$\frac{\partial Q}{\partial \beta_1} = 0$$

$$\Rightarrow \sum_{i=0}^{n} 2 \cdot (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-x_i) = 0$$

$$\Rightarrow -2 \cdot \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (x_i) = 0$$

$$\Rightarrow -\sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (x_i) = 0$$

$$\Rightarrow -\sum_{i=0}^{n} y_i \cdot x_i + \beta_0 \sum_{i=0}^{n} x_i + \beta_1 \cdot \sum_{i=0}^{n} x_i \cdot x_i = 0$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_1:

$$\Rightarrow -\sum_{i=0}^{n} y_i \cdot x_i + \beta_0 \sum_{i=0}^{n} x_i + \beta_1 \cdot \sum_{i=0}^{n} x_i \cdot x_i = 0$$

$$\Rightarrow -\sum_{i=0}^{n} y_i \cdot x_i + (\bar{y} - \beta_1 \cdot \bar{x}) \sum_{i=0}^{n} x_i + \beta_1 \cdot \sum_{i=0}^{n} x_i \cdot x_i = 0$$

$$\Rightarrow -\sum_{i=0}^{n} y_i \cdot x_i + \bar{y} \cdot \sum_{i=0}^{n} x_i - \beta_1 \cdot \bar{x} \sum_{i=0}^{n} x_i + \beta_1 \cdot \sum_{i=0}^{n} x_i^2 = 0$$

$$\Rightarrow -\beta_1 \cdot \bar{x} \sum_{i=0}^{n} x_i + \beta_1 \cdot \sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} y_i \cdot x_i - \bar{y} \cdot \sum_{i=0}^{n} x_i$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_1:

$$\Rightarrow \beta_1 \left[\sum_{i=0}^{n} x_i^2 - \sum_{i=0}^{n} x_i \cdot \bar{x}\right] = \left[\sum_{i=0}^{n} y_i \cdot x_i - \sum_{i=0}^{n} x_i \bar{y}\right]$$

$$\Rightarrow \beta_1 = \frac{\left[\sum_{i=0}^{n} y_i \cdot x_i - \sum_{i=0}^{n} x_i \bar{y}\right]}{\left[\sum_{i=0}^{n} x_i^2 - \sum_{i=0}^{n} x_i \cdot \bar{x}\right]}$$

$$\Rightarrow \beta_1 = \frac{\left[\sum_{i=0}^{n} y_i \cdot x_i\right] - n \cdot \left[\frac{\sum_{i=0}^{n} x_i}{n} \cdot \bar{y}\right]}{\left[\sum_{i=0}^{n} x_i^2\right] - n \cdot \left[\frac{\sum_{i=0}^{n} x_i}{n} \cdot \bar{x}\right]}$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_1:

$$\Rightarrow \beta_1 = \frac{\left[\sum_{i=0}^{n} y_i \cdot x_i\right] - n \cdot \left[\frac{\sum_{i=0}^{n} x_i}{n} \cdot \bar{y}\right]}{\left[\sum_{i=0}^{n} x_i^2\right] - n \cdot \left[\frac{\sum_{i=0}^{n} x_i}{n} \cdot \bar{x}\right]}$$

$$\Rightarrow \beta_1 = \frac{\left[\sum_{i=0}^{n} y_i \cdot x_i\right] - n \cdot [\bar{x} \cdot \bar{y}]}{\left[\sum_{i=0}^{n} x_i^2\right] - n \cdot [\bar{x} \cdot \bar{x}]}$$

Linear Regression

Solving the optimization problem The First Order Conditions w.r.t. β_1:

$$\Rightarrow \beta_1 = \frac{\frac{1}{n} \cdot \left[\sum_{i=0}^{n} y_i \cdot x_i\right] - \frac{1}{n} \cdot n \cdot [\bar{x} \cdot \bar{y}]}{\frac{1}{n} \cdot \left[\sum_{i=0}^{n} x_i^2\right] - \frac{1}{n} \cdot n \cdot [\bar{x} \cdot \bar{x}]}$$

$$\Rightarrow \beta_1 = \frac{E(xy) - E(x) \cdot E(y)}{E(x^2) - E(x) \cdot E(x)}$$

$$\Rightarrow \beta_1 = \frac{Cov(xy)}{Var(x)}$$

Linear Regression

The Correlation Coefficient

Besides the regression slope β_1 and intercept β_0, the third parameter of importance is the correlation coefficient r^2.

r^2 is the ratio between the variance in y and the variance that is "explained" by the regression line $\hat{y}$. Equivalently, it is the ratio of the variance in $\hat{y}$ to the total variance in y.

Linear Regression

The Correlation Coefficient

$$
\begin{aligned}
r^2 &= \frac{Var(\hat{y})}{Var(y)} \\
&= \frac{Var(\beta_0 + \beta_1 \cdot x)}{Var(y)} \\
&= \frac{\beta_1^2 \cdot Var(x)}{Var(y)} \\
&= \frac{Cov(xy)^2}{Var(x)^2} \cdot \frac{Var(x)}{Var(y)} \\
&= \frac{Cov(xy)^2}{Var(x)} \cdot \frac{1}{Var(y)}
\end{aligned}
$$

Linear Regression

The Correlation Coefficient Finally, the correlation coefficient, r is written as:

$$r = \sqrt{\frac{Cov(xy)}{Var(x) \cdot Var(y)}}$$

Linear Regression

The Standard Error

- Fourth parameter of interest is the standard error.
- From each observed value of x, the regression line gives a predicted value $\hat{y}$ that may differ from the observed value of y . This difference is the error of estimate can be given the symbol e.
- $e_i = y_i - \hat{y}_i$ for each observation i. Also note that ideally, $\sum e_i = 0$.

Linear Regression

The Standard Error

Just as deviations about the mean can be summarized into the standard deviation, so these errors can be summarized into a standard error.

The standard error of estimate is often given the symbol s_e and written as:

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

Linear Regression

Standard Error vs. Standard Deviation

- In the numerator, the deviations of the predicted values about the regression line, rather than the deviations of the observed values about the mean are used.
- The second difference is that the denominator is $n - 2$ rather than $n - 1$.
- In the case of s_e, this occurs because the deviations are about the regression line, and two values are required to fix a regression line.

Linear Regression *The Standard Error*

In order to carry out calculations in a timely manner, we will write the standard error in the following form

$$s_e = \sqrt{\frac{\sum y_i^2 - \beta_0 \cdot \sum y_i - \beta_1 \cdot \sum x_i \cdot y_i}{n-2}}$$

We are not gonna derive it for obvious reasons!

Linear Regression

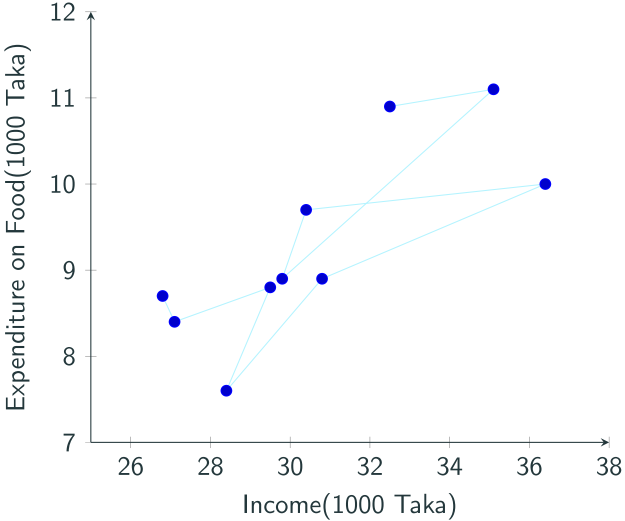
Example: Regression of Food Expenditure on Income

Let's model Income vs. Expenditure on Food, where the independent variable X is Income and the dependent variable Y is the Expenditure on Food.

We will find $Var(X)$, $Var(Y)$, $Cov(X, Y)$, β_0, β_1, R^2, s_e, s_b, carry out hypothesis testing and calculate the confidence interval.

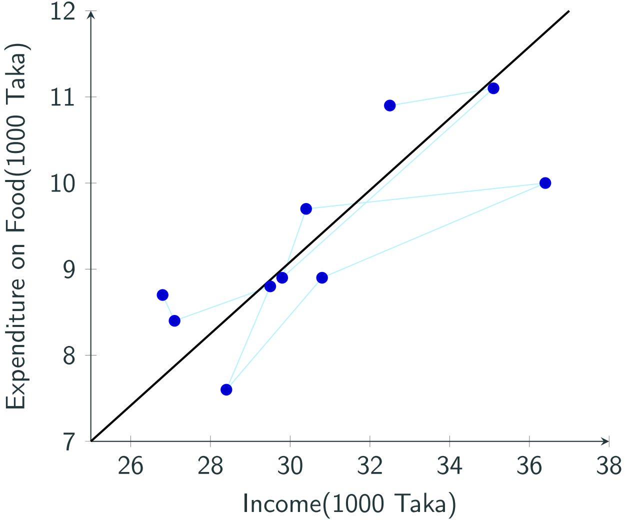
Linear Regression

An illustration



Linear Regression

An illustration



Linear Regression

X	Y	X^2	Y^2	$X \cdot Y$
26.8	8.7	718.24	75.69	233.16
27.1	8.4	734.41	70.56	227.64
29.5	8.8	870.34	77.44	259.60
28.4	7.6	806.56	57.76	215.84
30.8	8.9	948.64	79.21	274.12
36.4	10.0	1324.96	100	364.00
30.4	9.7	924.16	94.09	294.88
29.8	8.9	888.04	79.21	265.22
35.1	11.1	1232.01	123.21	389.61
32.5	10.9	1056.25	118.81	354.25
306.8	93.0	9503.52	875.98	2878.32

Linear Regression

First, calculate the following:

- $\sum X = 306.8$
- $\sum Y = 93.0$
- $\sum X^2 = 9503.52$
- $\sum Y^2 = 875.98$
- $\sum X \cdot Y = 2878.32$
- $n = 10$

Linear Regression

Calculate Covariance of X and Y

$$\begin{aligned} S_{XY} &= \sum X \cdot Y - \frac{(\sum X) \cdot (\sum Y)}{n} \\ &= 2878.32 - \frac{(306.8) \cdot (93.0)}{10} \\ &= 2878.32 - 2853.24 \\ &= 25.08 \end{aligned}$$

Linear Regression

Calculate Variance of X

$$
\begin{aligned}
S_{XX} &= \sum X^2 - \frac{(\sum X) \cdot (\sum X)}{n} \\
&= 9503.52 - \frac{(306.8)^2}{10} \\
&= 9503.52 - 9412.64 \\
&= 90.90
\end{aligned}
$$

Linear Regression

Calculate Variance of Y

$$
\begin{aligned}
S_{YY} &= \sum Y^2 - \frac{(\sum Y) \cdot (\sum Y)}{n} \\
&= 875.98 - \frac{(93.0)^2}{10} \\
&= 875.98 - 864.9 \\
&= 11.08
\end{aligned}
$$

Linear Regression

Calculate β_0 and β_1.

$$\begin{aligned} \beta_1 &= \frac{S_{XY}}{S_{XX}} = \frac{25.08}{90.90} = 0.276 \\ \beta_0 &= \bar{Y} - \beta_1 \cdot \bar{X} = \frac{93.0}{10} - 0.276 \cdot \frac{306.8}{10} \end{aligned} \quad (1)$$

Linear Regression

Calculate the Standard Error

$$
\begin{aligned}
s_e &= \sqrt{\frac{\sum Y^2 - \beta_0 \cdot \sum Y - \beta_1 \cdot \sum X \cdot Y}{n-2}} \\
&= \sqrt{\frac{875.98 - (0.832 \times 93.0) - (0.276 \times 2878.32)}{10-2}} \\
&= \sqrt{\frac{4.188}{8}} \\
&= \sqrt{0.523} \\
&= 0.724
\end{aligned}
$$

Linear Regression

Calculate the standard deviation of the sampling distribution of β_1

$$\begin{aligned} s_b &= \frac{s_e}{\sqrt{S_{XX}}} \\ &= \frac{0.724}{\sqrt{90.8}} \\ &= \frac{0.724}{9.534} \\ &= 0.0759 \end{aligned}$$

Linear Regression

Calculate R^2

$$
\begin{aligned}
R^2 &= \frac{S_{XY}^2}{(S_{XX} \cdot S_{YY})} \\
&= \frac{25.08^2}{90.90 \times 11.08} \\
&= \frac{629}{1007} \\
&= 0.625
\end{aligned}
$$

Linear Regression

Carry out Hypothesis Testing

$$\begin{aligned} H_0 &: x = 0 \\ H_1 &: x \neq 0 \end{aligned} \tag{2}$$

Linear Regression

Calculate t-statistic

$$
\begin{aligned}
t - statistic &= \frac{\beta_1 - x}{s_b} \\
&= \frac{(0.276 - 0)}{0.0756} \\
&= 3.651 \geq 3.355
\end{aligned}
$$

Notice that the p-value is less than 0.01. If we set $\alpha = 0.01$, we can conclude the following:

Since $p - value < \alpha$ we reject H_0. Therefore, at the 1 % level of significance there is evidence that there is a relationship between income and expenditure on food.

Linear Regression

Calculate 95% Confidence Interval, t-statistic = 2.306

Error Bound for the Mean

$$
\begin{aligned}
t_{\frac{\alpha}{2}} \times \frac{s_b}{\sqrt{n}} &= 2.306 \times \frac{0.0756}{\sqrt{10}} \\
&= 0.055
\end{aligned}
$$

Linear Regression

Calculate 95% Confidence Interval, t-statistic = 2.306

Error Bound for the Mean

$$\begin{aligned} 95\%\ CI &= (\beta_1 + EBM, \beta_1 - EBM) \\ &= (0.276 + 0.055, 0.276 - 0.055) \\ &= (0.331, 0.221) \end{aligned}$$

Linear Regression

This concludes our syllabus. Good luck with the final quiz!