Regression analysis

One of the most important kinds of optimization, especially throughout the economic, social, biological and physical sciences, is minimizing “how far our real-world data is from a functional model.”

Regression analysis

One of the most important kinds of optimization, especially throughout the economic, social, biological and physical sciences, is minimizing “how far our real-world data is from a functional model.” We will emphasize this kind of optimization, known as regression analysis, more than the book does.

Given any two data points, we can find a line through them.

Regression analysis

One of the most important kinds of optimization, especially throughout the economic, social, biological and physical sciences, is minimizing "how far our real-world data is from a functional model." We will emphasize this kind of optimization, known as regression analysis, more than the book does.

Given any two data points, we can find a line through them. We saw as an application of systems of linear equations, solved through matrices, that through any three points we can fit a parabola.

three points we can fit a parabola. This technique generalizes - through 100 points we can find a polynomial whose leading term is a multiple of x^{99}.

three points we can fit a parabola. This technique generalizes - through 100 points we can find a polynomial whose leading term is a multiple of x^{99}. But such a polynomial has 100 terms, and so is cumbersome as a functional model.

three points we can fit a parabola. This technique generalizes - through 100 points we can find a polynomial whose leading term is a multiple of x^{99}. But such a polynomial has 100 terms, and so is cumbersome as a functional model.

The simplest kind of functional model

three points we can fit a parabola. This technique generalizes - through 100 points we can find a polynomial whose leading term is a multiple of x^{99}. But such a polynomial has 100 terms, and so is cumbersome as a functional model.

The simplest kind of functional model - the one you always start with -

three points we can fit a parabola. This technique generalizes - through 100 points we can find a polynomial whose leading term is a multiple of x^{99}. But such a polynomial has 100 terms, and so is cumbersome as a functional model.

The simplest kind of functional model - the one you always start with - is a linear model. The idea of regression analysis is to find a line which does not have to go through the data points exactly, but which does do the best it can, minimizing the error involved as measured by (squares of) deviation in y values.

Definition 1. *The vertical deviation of a function* $f(x)$ *from some collection of data points* $\{(x_i, y_i)\}$ *is the sum*

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_i - f(x_i))^2 + \cdots .$$

Definition 1. *The vertical deviation of a function* $f(x)$ *from some collection of data points* $\{(x_i, y_i)\}$ *is the sum*

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_i - f(x_i))^2 + \cdots .$$

This definition is best pictured on the graph of the function and the data points.

Definition 1. *The vertical deviation of a function* $f(x)$ *from some collection of data points* $\{(x_i, y_i)\}$ *is the sum*

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_i - f(x_i))^2 + \cdots .$$

This definition is best pictured on the graph of the function and the data points. The reason that the differences are squares is so that the deviation is always positive and has a derivative which exists (which does not always happen when you make things positive by taking absolute value).

Definition 1. *The vertical deviation of a function* $f(x)$ *from some collection of data points* $\{(x_i, y_i)\}$ *is the sum*

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_i - f(x_i))^2 + \cdots .$$

This definition is best pictured on the graph of the function and the data points. The reason that the differences are squares is so that the deviation is always positive and has a derivative which exists (which does not always happen when you make things positive by taking absolute value).

Example 2. *Find the line whose vertical deviation from the points* $(1,1)$, $(2,3)$ *and* $(3,4)$ *is minimal.*

Example 2. *Find the line whose vertical deviation from the points* $(1,1)$, $(2,3)$ *and* $(3,4)$ *is minimal.*

Some important features of this example:

Example 2. *Find the line whose vertical deviation from the points* $(1,1)$, $(2,3)$ *and* $(3,4)$ *is minimal.*

Some important features of this example:

- Though it looks like x and y should be our variables, the slope m and y-intercept b of the line are the "real" variables.

Example 2. *Find the line whose vertical deviation from the points* $(1, 1)$, $(2, 3)$ *and* $(3, 4)$ *is minimal.*

Some important features of this example:

- Though it looks like x and y should be our variables, the slope m and y-intercept b of the line are the "real" variables. Much as the coefficients of a quadratic polynomial are variables when we fit a parabola to data.

Example 2. *Find the line whose vertical deviation from the points* $(1, 1)$, $(2, 3)$ *and* $(3, 4)$ *is minimal.*

Some important features of this example:

- Though it looks like x and y should be our variables, the slope m and y-intercept b of the line are the "real" variables. Much as the coefficients of a quadratic polynomial are variables when we fit a parabola to data.

- Ultimately, the critical point is found as a solution of a system of linear equations.

Definition 3. *The linear regression line for a collection of data is the linear function whose vertical deviation from that collection is minimal.*

The ability to find such lines is programmed into statistical and data analysis software, as well as your calculators.

Definition 3. *The linear regression line for a collection of data is the linear function whose vertical deviation from that collection is minimal.*

The ability to find such lines is programmed into statistical and data analysis software, as well as your calculators. The book gives explicit formulae which you may use for the homework, but for the exam it will be more important that you understand the way in which minimization techniques are employed.

Definition 3. *The linear regression line for a collection of data is the linear function whose vertical deviation from that collection is minimal.*

The ability to find such lines is programmed into statistical and data analysis software, as well as your calculators. The book gives explicit formulae which you may use for the homework, but for the exam it will be more important that you understand the way in which minimization techniques are employed. (This is another case where we are learning exactly what our calculators are doing behind the scenes).

Theorem 4. *The linear regression line for the collection of data* (x_1, y_1), $(x_2, y_2), \ldots, (x_k, y_k)$ *is the function* $f(x) = mx + b$ *such that the sum*

$$(x_1 m + b - y_1)^2 + \cdots + (x_k m + b - y_k)^2$$

is minimized.

Theorem 4. *The linear regression line for the collection of data* (x_1, y_1), $(x_2, y_2), \ldots, (x_k, y_k)$ *is the function* $f(x) = mx + b$ *such that the sum*

$$(x_1 m + b - y_1)^2 + \cdots + (x_k m + b - y_k)^2$$

is minimized. Taking the partial derivatives and setting them to zero leads to a system of two linear equations in the variables m *and* b.

Theorem 4. *The linear regression line for the collection of data* (x_1, y_1), $(x_2, y_2), \ldots, (x_k, y_k)$ *is the function* $f(x) = mx + b$ *such that the sum*

$$(x_1 m + b - y_1)^2 + \cdots + (x_k m + b - y_k)^2$$

is minimized. Taking the partial derivatives and setting them to zero leads to a system of two linear equations in the variables m *and* b.

One of the main applications of linear regression is to fill in/ predict/ extrapolate values for a function from known values.

Theorem 4. *The linear regression line for the collection of data* (x_1, y_1), $(x_2, y_2), \ldots, (x_k, y_k)$ *is the function* $f(x) = mx + b$ *such that the sum*

$$(x_1 m + b - y_1)^2 + \cdots + (x_k m + b - y_k)^2$$

is minimized. Taking the partial derivatives and setting them to zero leads to a system of two linear equations in the variables m *and* b*.*

One of the main applications of linear regression is to fill in/ predict/ extrapolate values for a function from known values.

Example 5. *Because of a computer error, some of the sales figures for a real estate company were lost. The sales figures (measured in millions of dollars) which are available for Gary Gladhand are:*

Example 5. *Because of a computer error, some of the sales figures for a real estate company were lost. The sales figures (measured in millions of dollars) which are available for Gary Gladhand are:*

1998	*1999*	*2001*	*2003*
0.9	*1.5*	*1.9*	*2.4*

Example 5. *Because of a computer error, some of the sales figures for a real estate company were lost. The sales figures (measured in millions of dollars) which are available for Gary Gladhand are:*

1998	*1999*	*2001*	*2003*
0.9	*1.5*	*1.9*	*2.4*

Graph these, and then use the regression line to estimate/predict what he sold in 2000 and what he will sell in 2004.

Example 5. *Because of a computer error, some of the sales figures for a real estate company were lost. The sales figures (measured in millions of dollars) which are available for Gary Gladhand are:*

1998	1999	2001	2003
0.9	1.5	1.9	2.4

Graph these, and then use the regression line to estimate/predict what he sold in 2000 and what he will sell in 2004.

We see that as predicted by the theorem, if we follow the steps outlined for finding and classifying critical points, the equations we solve for the slope m and y-intercept b are always, at the end of the day, simple linear equations.

Important warning:

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions.

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions. For example, if you take points from the graph of an exponential function a linear fit will always under-estimate in the long term.

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions. For example, if you take points from the graph of an exponential function a linear fit will always under-estimate in the long term. One can instead do regression with other kinds of functions.

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions. For example, if you take points from the graph of an exponential function a linear fit will always under-estimate in the long term. One can instead do regression with other kinds of functions. The way to fit an exponential function is to take the logarithm of the data and fit that to a linear function.

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions. For example, if you take points from the graph of an exponential function a linear fit will always under-estimate in the long term. One can instead do regression with other kinds of functions. The way to fit an exponential function is to take the logarithm of the data and fit that to a linear function.

Example 6. *Suppose that an investment had yearly values of*

1999	*2000*	*2001*	*2002*	*2003*
48K	*52K*	*58K*	*76K*	*96K*

Important warning: If the data is being sampled is not close to linear, then linear regression can produce spectacularly incorrect predictions. For example, if you take points from the graph of an exponential function a linear fit will always under-estimate in the long term. One can instead do regression with other kinds of functions. The way to fit an exponential function is to take the logarithm of the data and fit that to a linear function.

Example 6. *Suppose that an investment had yearly values of*

1999	*2000*	*2001*	*2002*	*2003*
48K	*52K*	*58K*	*76K*	*96K*

Use linear

regression analysis, after taking logarithms, to find the function of the form Ae^{rt} *which best fits these values, and thus deduce the growth rate.*

regression analysis, after taking logarithms, to find the function of the form Ae^{rt} *which best fits these values, and thus deduce the growth rate.*