Linear Regression

Linear regression is a statistical technique often used to predict one variable from another.

Linear Regression

Linear regression is a statistical technique often used to predict one variable from another.

Usually the basis for this is a *linear association* between two variables.

Linear Regression

Linear regression is a statistical technique often used to predict one variable from another.

Usually the basis for this is a *linear association* between two variables.

Usually this association is assumed to have the following form:

$$y = mx + b + e$$

Linear Regression

Linear regression is a statistical technique often used to predict one variable from another.

Usually the basis for this is a *linear association* between two variables.

Usually this association is assumed to have the following form:

$$y = mx + b + e$$

- x is the predictor variable
- y is the dependent or predicted variable
- m is the slope of the regression line
- b is the intercept of the regression
- e has a bell curve distribution with mean zero

Linear Regression

The *correlation coefficient* r is a measure of *linear* assiciation between two variables.

Linear Regression

The *correlation coefficient* r is a measure of *linear* assiciation between two variables.

An r value of 1 or -1 indicates a perfect linear relationship, $y = mx + b$

Linear Regression

The *correlation coefficient* r is a measure of *linear* assiciation between two variables.

An r value of 1 or -1 indicates a perfect linear relationship, $y = mx + b$

An r value of 0 indicates no linear relationship.

Linear Regression

The *correlation coefficient* r is a measure of *linear* assiciation between two variables.

An r value of 1 or -1 indicates a perfect linear relationship, $y = mx + b$

An r value of 0 indicates no linear relationship.

This is equivalent to saying that the slope m is zero.

Linear Regression

The slope m, correlation coefficient r, and the standard deviations SD_x and SD_y are related by:

$$m = \frac{r \cdot SD_y}{SD_x}$$

Linear Regression

The slope m, correlation coefficient r, and the standard deviations SD_x and SD_y are related by:

$$m = \frac{r \cdot SD_y}{SD_x}$$

Notice that m is necessarily zero if r is zero:

$$m = \frac{0 \cdot SD_y}{SD_x}$$

so

$$m = 0$$

Linear Regression

The slope m, intercept b, and the means $\overline{x}, \overline{y}$ are related by:

$$b = \overline{y} - m \cdot \overline{x}$$

Linear Regression

The slope m, intercept b, and the means $\overline{x}, \overline{y}$ are related by:

$$b = \overline{y} - m \cdot \overline{x}$$

Note that if the slope is zero, the intercept is $\overline{y}$

$$b = \overline{y} - 0 \cdot \overline{x}$$

so

$$b = \overline{y}$$

Linear Regression

The slope m, intercept b, and the means $\overline{x}, \overline{y}$ are related by:

$$b = \overline{y} - m \cdot \overline{x}$$

Linear Regression

The slope m, intercept b, and the means $\overline{x}, \overline{y}$ are related by:

$$b = \overline{y} - m \cdot \overline{x}$$

Note that if the slope is zero, the intercept is $\overline{y}$

$$b = \overline{y} - 0 \cdot \overline{x}$$

so

$$b = \overline{y}$$

Linear Regression

The **RMS error** s is a measure of the distance from the regression line.

Linear Regression

The **RMS error** s is a measure of the distance from the regression line.

The term e in the usual linear model

$$y = mx + b + e$$

is assumed to have a bell curve distribution with mean zero.

Linear Regression

The **RMS error** s is a measure of the distance from the regression line.

The term e in the usual linear model

$$y = mx + b + e$$

is assumed to have a bell curve distribution with mean zero.

The standard deviation of this bell curve is the RMS error, s.

Linear Regression

The **RMS error** has characteristics similar to the standard deviation for a bell curve.

Linear Regression

The **RMS error** has characteristics similar to the standard deviation for a bell curve.

If we take a scatter plot and draw the regression line on it,

68% of the observations will fall in a band of width s on either side of the regression line.

Linear Regression

The **RMS error** has characteristics similar to the standard deviation for a bell curve.

If we take a scatter plot and draw the regression line on it,

68% of the observations will fall in a band of width s on either side of the regression line.

About 95% will fall in a band of with $2s$ on either side of the regression line.

Linear Regression

The **RMS error** is given by the formula:

$$s = \sqrt{1 - r^2} \cdot SD_y$$

where r is the correlation coefficient.

Linear Regression

The **RMS error** is given by the formula:

$$s = \sqrt{1 - r^2} \cdot SD_y$$

where r is the correlation coefficient.

The closer r is to -1 or 1, the smaller the RMS error becomes.

Linear Regression

The square of the correlation coefficient r is used so often it is given a name.

Linear Regression

The square of the correlation coefficient r is used so often it is given a name.

Definition: The **coefficient of determination**, denoted by R^2, is the square of the correlation coefficient r.

Linear Regression

The square of the correlation coefficient r is used so often it is given a name.

Definition: The **coefficient of determination**, denoted by R^2, is the square of the correlation coefficient r.

$$R^2 = r^2$$

Linear Regression

The square of the correlation coefficient r is used so often it is given a name.

Definition: The **coefficient of determination**, denoted by R^2, is the square of the correlation coefficient r.

$$R^2 = r^2$$

The coefficient of determination R^2 can be interpreted as the *proportion of the dispersion of y that is explained by the regression line*.

Linear Regression

Most spreadsheets have a function called CORREL that will calculate a correlation coefficient r

Linear Regression

Most spreadsheets have a function called CORREL that will calculate a correlation coefficient r

For example, if there are 40 pairs of x and y values in columns A and B,

=CORREL(A1:A40,B1:B40)

will compute the correlation coefficient r.

Linear Regression

Most spreadsheets have a function called CORREL that will calculate a correlation coefficient r

For example, if there are 40 pairs of x and y values in columns A and B,

=CORREL(A1:A40,B1:B40)

will compute the correlation coefficient r.

The exact name and syntax of this function will vary somewhat among the different brands of spreadsheet programs.

Linear Regression

Most spreadsheets have a functions called SLOPE and INTERCEPT that will calculate m and b for a regression line.

Linear Regression

Most spreadsheets have a functions called SLOPE and INTERCEPT that will calculate m and b for a regression line.

For example, if there are 40 pairs of x and y values in columns A and B,

=SLOPE(A1:A40,B1:B40)

will compute the slope m.

Linear Regression

Most spreadsheets have a functions called SLOPE and INTERCEPT that will calculate m and b for a regression line.

For example, if there are 40 pairs of x and y values in columns A and B,

=SLOPE(A1:A40,B1:B40)

will compute the slope m.

=INTERCEPT(A1:A40,B1:B40)

will compute the intercept b.

Linear Regression

The RMS error s can be computed as

$$s = \sqrt{1 - r^2} \cdot SD_y$$

Linear Regression

The RMS error s can be computed as

$$s = \sqrt{1 - r^2} \cdot SD_y$$

If there are 40 pairs of x and y values in columns A and B,

$= SQRT(1 - (CORREL(A1 : A40, B1 :$
$B40)^2) * STDEV(B1 : B40)$

will compute the RMS error s.

Linear Regression

One of the most common uses of regression is to estimate the rate of growth of some quantity.

Linear Regression

One of the most common uses of regression is to estimate the rate of growth of some quantity.

In this type of application, the x values represent time.

Linear Regression

One of the most common uses of regression is to estimate the rate of growth of some quantity.

In this type of application, the x values represent time.

The y values represent the quantity we want to determine the growth rate of.

The slope represents the increase in the quantity measured per unit of time.

Linear Regression

The regression models we have studied so far will work fine in this situation, provided the following assumption is reasonable:

Linear Regression

The regression models we have studied so far will work fine in this situation, provided the following assumption is reasonable:

The change in y measured in *units* is the same, on average, for each unit of time.

Linear Regression

The regression models we have studied so far will work fine in this situation, provided the following assumption is reasonable:

The change in y measured in *units* is the same, on average, for each unit of time.

For example, if we are measuring cars produced, we can assume that the **number** of cars produced increases or decreases by the same amount each month.

Linear Regression

The regression models we have studied so far will work fine in this situation, provided the following assumption is reasonable:

The change in y measured in *units* is the same, on average, for each unit of time.

For example, if we are measuring cars produced, we can assume that the **number** of cars produced increases or decreases by the same amount each month.

That number is the slope of the regression line.

Linear Regression

In many situations, especially in business, a different assumption is made about the nature of growth.

Linear Regression

In many situations, especially in business, a different assumption is made about the nature of growth.

In these applications, it is assumed that the *percentage* change in y from month to month is constant.

Linear Regression

In many situations, especially in business, a different assumption is made about the nature of growth.

In these applications, it is assumed that the *percentage* change in y from month to month is constant.

This creates a problem, because x and y *no longer have a linear relationship*

Linear Regression

In many situations, especially in business, a different assumption is made about the nature of growth.

In these applications, it is assumed that the *percentage* change in y from month to month is constant.

This creates a problem, because x and y *no longer have a linear relationship*

That is, the equation

$$y = mx + b + e$$

no longer holds.

Linear Regression

In a constant percentage growth situation, if we plot y and x over time, we *do not* get a straight line:

$$y = mx + b$$

Linear Regression

In a constant percentage growth situation, if we plot y and x over time, we *do not* get a straight line:

$$y = mx + b$$

Constant percentage growth produces an *exponential curve*. One formulation is:

$$y = b \cdot m^x$$

Linear Regression

In a constant percentage growth situation, if we plot y and x over time, we *do not* get a straight line:

$$y = mx + b$$

Constant percentage growth produces an *exponential curve*. One formulation is:

$$y = b \cdot m^x$$

Generally speaking, a curve is much more difficult to fit to data than a straight line.

Linear Regression

One way to handle this is with a *transformation*

Linear Regression

One way to handle this is with a *transformation*

In general a transformation turns data that fits one model into data that fits another.

Linear Regression

One way to handle this is with a *transformation*

In general a transformation turns data that fits one model into data that fits another.

Hopefully, the second model is easier to work with.

Linear Regression

One way to handle this is with a *transformation*

In general a transformation turns data that fits one model into data that fits another.

Hopefully, the second model is easier to work with.

Once we have the fitted or projected values, we reverse the transformation to recover the original measures.

Linear Regression

There are many transformations, but the one that works in this case is the *log transform*

Linear Regression

There are many transformations, but the one that works in this case is the *log transform*

If we take (natural) logs of both sides of the equation

$$y = b \cdot m^x$$

we get

$$\ln y = \ln b + \ln m \cdot x$$

Linear Regression

There are many transformations, but the one that works in this case is the *log transform*

If we take (natural) logs of both sides of the equation

$$y = b \cdot m^x$$

we get

$$\ln y = \ln b + \ln m \cdot x$$

Now we have a linear equation instead of an exponential one.

Linear Regression

The inverse of the log transform is the *exponential*, usually denoted by EXP

Linear Regression

The inverse of the log transform is the *exponential*, usually denoted by EXP

To get back to a model for the untransformed data, we apply the inverse of the transform to the fitted y values, the slope, and the intercept. For the original model,

- $m = EXP(SLOPE)$
- $b = EXP(INTERCEPT)$
- $y = EXP(SLOPE * x + INTERCEPT)$

Linear Regression

The inverse of the log transform is the *exponential*, usually denoted by EXP

To get back to a model for the untransformed data, we apply the inverse of the transform to the fitted y values, the slope, and the intercept. For the original model,

- $m = EXP(SLOPE)$
- $b = EXP(INTERCEPT)$
- $y = EXP(SLOPE * x + INTERCEPT)$

Use these values with the model

$$y = b \cdot m^x$$