Topic 1: Simple Regression Analysis

Note Title 11/9/2006

- In many situations we are not just trying to estimate a parameter.
- Rather, we are interested in estimating the relationship among parameters.

Example: Suppose you are interested in estimating the relationship between heights and weights of Rutgers students. Specifically, suppose you believe that the following linear relationship exists:

$$\text{weight} = b_0 + b_1 \times \text{height}$$

What we don't know here are the multipliers or coefficients b_0 and b_1. We have a sample of say 50 Rutgers students and we tabulate their weights and heights.

student	weight	height
1	150 lb	5'4"
2	180 lb	5' 10"
⋮	⋮	
50	212 lb	6' 3"

From this data we are to estimate the relationship between height and weight. If we postulate that height and weight are related by the formula $\text{weight} = b_0 + b_1 \text{height}$, then all we need to do here is estimate b_0 and b_1.
In fact b_0 and b_1 are parameters of the entire Rutgers students population. Any estimate $\hat{b}_0$ and $\hat{b}_1$ are statistics, which have to be calculated from the sample.

Example 2: Suppose we like to estimate the relationship between selling price of a house in a large suburban

neighborhood and the square foot, unemployment rate in the county the neighborhood resides in, and the property tax rate.

Let us define:

Y = selling price

X_1 = square feotage

X_2 = unemployment rate

X_3 = property tax rate

Suppose we have modeled the following formula:

$$Y = f(X_1, X_2, X_3) = b_0 + b_1 X_1^2 + b_2 X_2 + b_3 X_3 + b_4 X_2 X_3$$

The functional relationship $f(X_1, X_2, X_3)$ has a formula, but the coefficients b_0, b_1, b_2 b_3 and b_4 are unknown. Thus to determine this relationship we need to estimate the coefficients b_0, b_1, b_2, b_3, b_4.

Note: In this example

- we have several variables: X_1, X_2, X_3
- The relationship $f(X_1, X_2, X_3)$ is not linear in X_1, X_2, & X_3.

Important → • However, the unknown parameters b_0, b_1, b_2, b_3 and b_4 are linear!

The Regression Model: In the regression model we have the following ingredients:

- The dependent or response variable (Y)

  This is the quantity whose behavior depends on other variables.

  - In example 1 weight is the response or dependent variable.
  - In example 2 sale price is the response or dependent variable.

Although in more sophisticated models, we may have several response variables, in our course we consider only models with one response variable

- The independent or explanatory or predictor variables ($X_1, X_2, \ldots, X_p$)
  These are variables whose values influence the value of the response variable:
  - In example 1 height is the independent or explanatory variable. Intuitively, the taller a person, the heavier we expect them to be.
  - In example 2 square footage, unemployment rate and local property taxes are explanatory or independent variables

In general we may have one or more dependent variables.

- The model or formula ($Y = f(X_1, \ldots, X_p; b_0, b_1, \ldots, b_m)$)
  This is the mathematical formula or relationship that connects the explanatory variables $X_1, X_2, \ldots, X_m$ to the response variable Y.

- In example 1, the formula $Y = f(X; b_0, b_1)$

weight $= b_0 + b_1 *$ height

- In example 2 the formula: $Y = f(X_1, X_2, X_3; b_0, b_1, b_2, b_3, b_4)$ is given explicitly by:

$$Y = b_0 + b_1 X_1^2 + b_2 X_2 + b_3 X_3 + b_4 X_2 X_3$$

- Unknown parameters $(b_0, b_1, \ldots, b_m)$

Our formula may express the relationship between response variables and explanatory variables. But this relationship may contain constants that are multiplied or added to make the formula. These constants are the unknown parameters

- These parameters must be estimated from the sample.
- Suppose $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$ are the estimates then the estimated formula: $\hat{y} = f(x_1, \ldots, x_p, \hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m)$

Example 1: the estimated b_0, and b_1 may turn out to be $\hat{b}_0$ and $\hat{b}_1$. For example it may turn out that $\hat{b}_0 = 20$ $\hat{b}_1 = 30$, then the estimated formula would be

$$\text{weight} = 20 + 30 * \text{height}$$

Now this formula can be used to estimate the expected weight of a 6' 3" person: 6' 3" = 6.25 so

$$\hat{y} = E(\text{weight}) = 20 + 30 \times 6.25 = 207.5 \text{ lb.}$$

In the remaining lectures, we will study how to

- estimate the parameter $b_0, \dots, b_m$
- How to calculate confidence intervals for each b_i
- How to calculate the estimated response value $\hat{y}$ and find confidence intervals for it

- How to evaluate the relevance of explanatory variables (they may not have any "explanatory power" at all.)
- How to evaluate the overall quality of the model.

Simple Linear Model

The simplest form of linear models is when we have only one explanatory variable, X, and one response variable Y and the assumed relationship is linear:

$$Y = b_0 + b_1 X$$

The underlying model and assumptions are as follows:

1) The data are presented as follows:

Y	X
Y_1	X_1
Y_2	X_2
$\vdots$	$\vdots$
Y_n	X_n

2)

- It is assumed that the $x_1, \ldots, x_n$ are designed by the experimenter.
- the values $Y_1, \ldots, Y_n$ are observed values
- the Y_i's are assumed to equal:

$$Y_i = b_0 + b_1 x_i + \epsilon_i$$

Y_i: observed; b_0, b_1: unknown Parameters; x_i: designed; ϵ_i: a random error

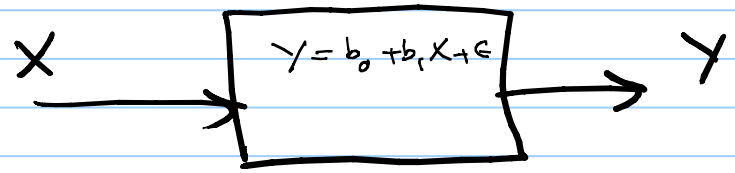
- The random error ϵ_i is what makes the model probabilistic not deterministic.

Note: If there was no random error ϵ_i, that is if the model was deterministic, then we would have needed just Two independent and two observations.

$Y_1 = b_0 + b_1 X_1$
$Y_2 = b_0 + b_1 X_2$
} This is a system of two linear equations and two unknowns: b_0 and b_1

The unpredictable random errors ϵ_i make it necessary to use all observations.

- The model can be visualized by a black box model:



The experimenter inserts values of $X_1, X_2 \ldots, X_n$, and observes values of $Y_1, Y_2 \ldots Y_n$.

- It is possible to insert the same value of x, that is it is possible that $X_1 = X_2$, but $Y_1 \neq Y_2$, because we may have different random errors.

Example Suppose we wish to use X = height as the explantory variable for Y = weight, for Rutgers students.

- As experimenters we select random students with given height and then observe their weights. Our design may be:

  1) - select 5 students at 5'0"
     - 5 students at 5'1"
     - 5 student at 5'2"
     - ⋮
     - 5 students at 6'10"

Clearly not all 5'0" students will have the same weight

Note: In this model $X_1, X_2 \ldots, X_n$ are fixed numbers determined by the experimenter

But $Y_1, Y_2 \ldots, Y_n$ are random variables

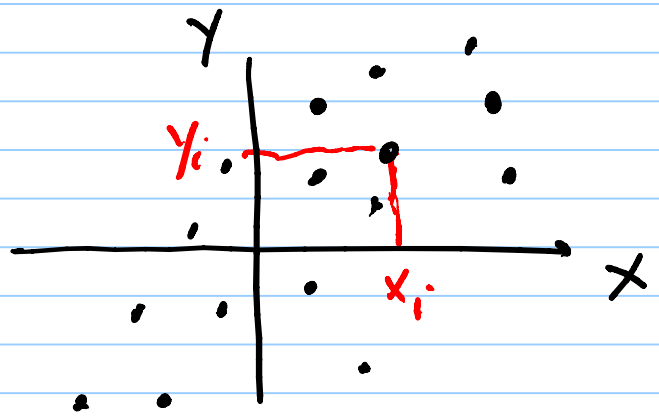
and in fact statistics because the random error ϵ_i in $Y_i = b_0 + b_1 X_i + \epsilon_i$ makes their values random.

An alternative approach to example 1.

- Suppose instead of selecting students of given heights, we select a random sample of students and record both their height ($= X$) and their weight ($= Y$). In this case both X and Y are random variables.
- this approach is not a designed approach and its analysis is more complicated.
- However the analysis for designed approach (i.e. X fixed) can be applied if distributions are conditioned (using conditional probabilities) but this is outside of our course.

Graphical representation of regression data

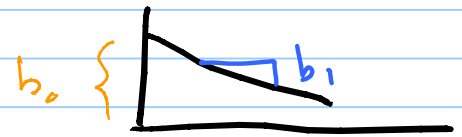
- In simple regression data come in the form of pairs (x,y)
- This can be represented by points on x-y plane.

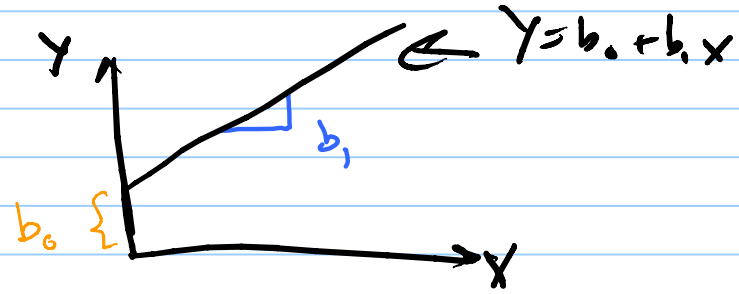


- In excel choose graphics button and select "scatter plot"

Estimating the parameters b_0 and b_1

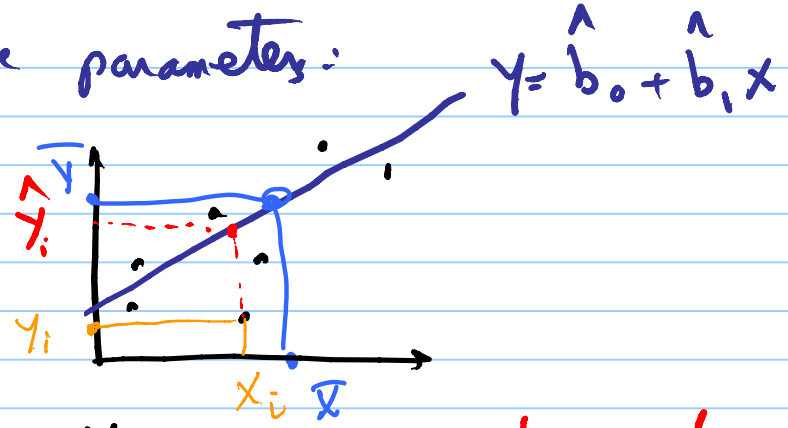
- b_0 is called intercept
- b_1 is called slope





Least squares estimate of the parameters:

- If the estimated relationship between x and y is $\hat{y} = \hat{b}_0 + \hat{b}_1 x$ then $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ is where the estimated y is while y_i is the observed y.



The difference $y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i$ is the observed error or residual at x_i.

- The least squares estimates of b_0 and b_1 are obtained by minimizing sum of squares of residuals:

Find $\hat{b}_0$ and $\hat{b}_1$ such that

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{b}_0 - \hat{b}_1 x_i)^2$$

is minimized

$\hat{\epsilon}_i$

Note Taking derivatives with respect to $\hat{b}_0$ and $\hat{b}_1$ results in a linear system of two equations in two unknowns $\hat{b}_0$ and $\hat{b}_1$.

- If

$$S_x = x_1 + \cdots + x_n \qquad S_y = y_1 + \cdots + y_n$$

$$S_{xx} = x_1^2 + \cdots + x_n^2 \qquad S_{yy} = y_1^2 + \cdots + y_n^2$$

$$S_{xy} = x_1 y_1 + \cdots + x_n y_n$$

then

$$\hat{b}_1 = \frac{n S_{xy} - S_x \cdot S_y}{n S_{xx} - S_x^2} \qquad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

(In excel $\hat{b}_0$ = INTERCEPT(.,.)

$\hat{b}_1$ = SLOPE(.,.))

Another convenient formula to describe the estimated coefficients:

Define: $SS_x = \sum_i (x_i - \bar{x})^2$ in EXCEL: VAR(x1:xn)*(n-1)

$SS_y = \sum_i (y_i - \bar{y})^2$ in EXCEL: VAR(Y1:Yn)*(n-1)

$SS_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$

in EXCEL: COVAR(Y1:Yn, X1:Xn)*n

then $\hat{b}_1 = \frac{SS_{xy}}{SS_x}$ in EXCEL: SLOPE(Y1:Yn, X1:Xn)

$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$ in EXCEL: INTERCEPT(Y1:Yn, X1:Xn)

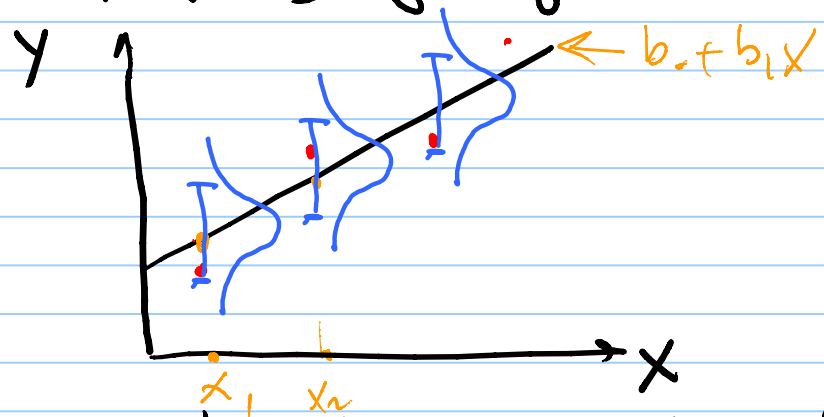
The probabilistic model for error:

We assume that the random errors ϵ_i satisfy the following properties:

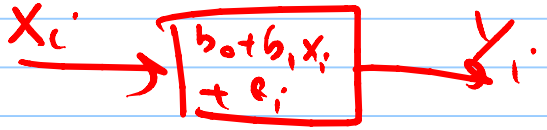
1) The ϵ_i independent: The error at x_i is independent of error at x_j

2) The ϵ_i are normal with common (but generally unknown) variance σ^2 and mean 0: $\epsilon_i \sim N(0, \sigma)$

The property of equal variance is called **homoscedasticity**



$$Y_i = b_0 + b_1 X_i + \epsilon_i$$



3) This implies that the observed values: $Y_i = b_0 + b_1 X_i + \epsilon_i$ are also normal $Y_i \sim N(b_0 + b_1 x_i, \sigma)$

4) The distribution of the statistics $\hat{b}_0$ and $\hat{b}_1$ are

also normal. with mean b_0 and b_1 (the true intercept and and slope) and standard deviation:

$$\hat{b}_0 \sim N(b_0, \sigma(b_0)) \qquad \hat{b}_1 \sim N(b_1, \sigma(b_1))$$

$S_{xx} = \sum (x_i - \bar{x})^2$

$$\sigma(\hat{b}_0) = \sigma\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right) \qquad \sigma(\hat{b}_1) = \frac{\sigma}{\sqrt{SS_x}}$$

5) The estimate of σ^2 is $\frac{SSE}{n-2} = \frac{\sum (\hat{y}_i - y_i)^2}{n-2}$

The quantity

$$SE = \sqrt{\frac{SSE}{n-2}}$$ is called **standard error**.

And the quantities $\hat{y}_i - y_i$ are called **residuals**

Note: Always $\sum$ residuals $= \boxed{\sum_i (\hat{y}_i - y_i) = 0}$

6) Defining $s(\hat{b}_0) = SE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)$ $\hat{s}(b_1) = \frac{SE}{\sqrt{S_{xx}}}$

as estimates of $\sigma(\hat{b}_0)$ and $\sigma(\hat{b}_1)$ we have that

$$\frac{\hat{b}_0 - b_0}{s(\hat{b}_0)} \sim t(n-2) \qquad \frac{\hat{b}_1 - b_1}{s(\hat{b}_1)} \sim t(n-2)$$

Testing the hypothesis that X is relevant in determining Y

How do we determine if the explanatory variable X actually determines, at least in part, the response variable Y?

Observe: If in the model $\hat{y} = b_0 + b_1 X$

$b_1 = 0$ then this is the same as saying X is irrelevant to Y

Thus to see if X is relevant we test whether $b_1 = 0$ or not.

Null hypothesis H_0: X is irrelevant in determining Y ($b_1 = 0$)

Alt hypothesis H_1: X is relevant in determining Y ($b_1 \neq 0$)

$$t\text{-value} = \frac{\hat{b}_i - 0}{s(b_i)}$$

(hypothesized value)

$$p\text{-value} = TDIST(t\text{-value}, n-2)$$

Confidence interval for the b_i:

$$h = TINV(\alpha, n-2)$$

$$\text{Half width} = s(b_i)\,h$$

with probability $1-\alpha$

$$\hat{b}_i - \text{Halfwidth} \leq b_i \leq \hat{b}_i + \text{Halfwidth}$$

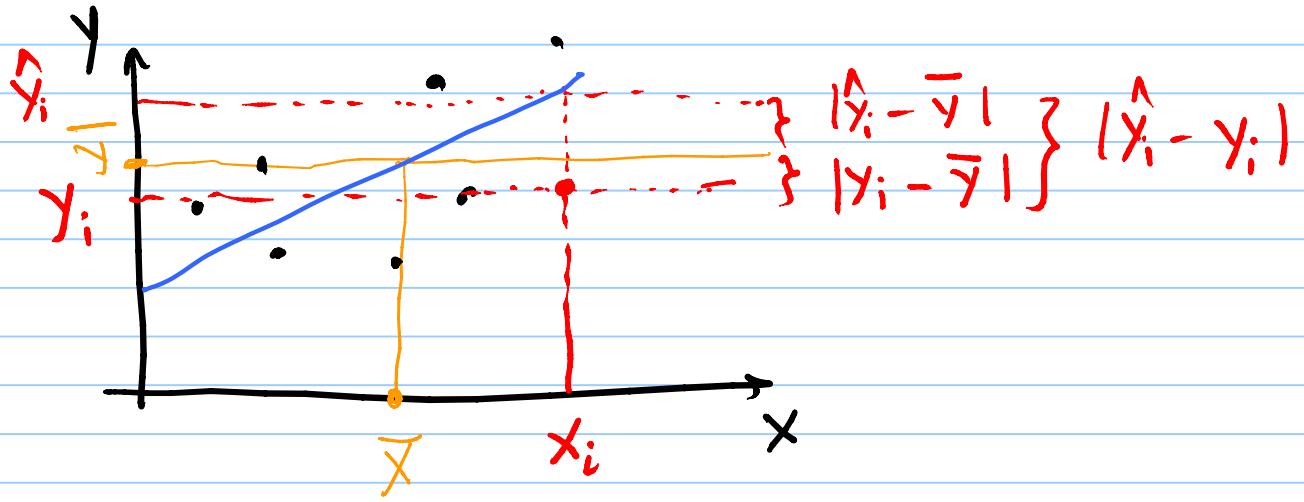
Note: As usual, since the hypothesis testing is 2-tail, if the confidence interval covers 0, we cannot reject the null hypothesis (that $b_i = 0$)

Note:

In Excel all these values are calculated for you by Tools → data analysis → Regression

Sums of Squares, R^2 and R:

There are three very important and "sums of squares" quantities associated with regression analysis:



- The quantity $Y_i - \bar{Y}$ is the deviation of the observed quantity Y_i from the average of all observations
- The quantity $Y_i - \hat{Y}_i$ is the deviation of observed quantity Y_i from what the estimated equation

predicts Y_i is supposed to be : $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$

- Finally, the quantity $\hat{Y}_i - \bar{Y}$ is the deviation of predicted value $\hat{Y}_i$ from the average of all observed Y_i, i.e. $\bar{Y}$

Notice that for X_i in this graph, obviously,

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

What is interesting is that squaring and adding these quantities:

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

Total variation

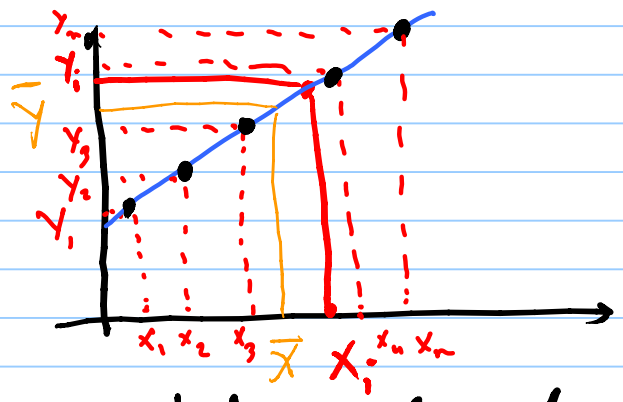
variation due to prediction error (residuals)

variation due to regression (changes in X_i's)

To understand what the meaning of this equality is, consider two extreme cases:

Case 1 Perfect fit:

Consider the following situation:



When we have all our data lined up perfectly, the quantities $\hat{y}_i$ and y_i (observed) match. Thus,

$$\hat{y}_i - y_i = 0 \quad \text{for each } i$$

This means that in this situation

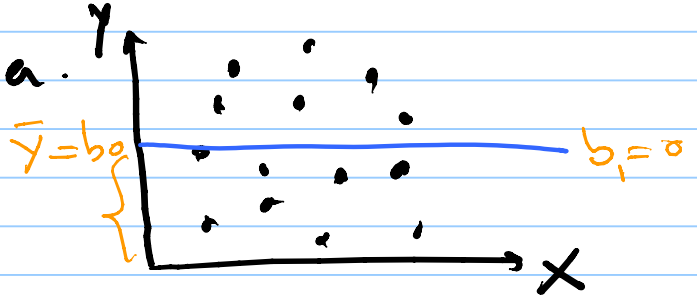
$$SST = \underbrace{SSE}_{=0} + SSR$$

and thus $SST = SSR$.

Thus, in this case the total deviation is entirely due to regression.

Case 2: complete irrelevance

Now consider a situation in which X is completely irrelevant to Y. In that case we know that in the equation $Y = b_0 + b_1 X$, $b_1 = 0$. Also in this case $b_0 = \bar{Y}$ (because $\bar{y} = b_0 + b_1 \bar{x}$ must be satisfied.) Thus each $\hat{y}_i = \bar{y}$ and thus $SSR = \sum (\hat{Y}_i - \bar{Y})^2 = 0$ and therefore $SST = SSE$. In other words the variation is completely due to error terms ϵ_i; the equation itself has no explanatory power.



The quantity

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

is called the coefficient of determination.

- When the data is perfectly lined up, $SST = SSR$ and $R^2 = 1$
- When the X is completely irrelevant to Y and $SST = SSE$ then $R^2 = 0$
- In general $0 \leq R^2 \leq 1$ the closer R^2 is to 1 the better the equation $Y = b_0 + b_1 X$ explains the the behavior of Y as a function of X.
- Interpretation of R^2:

  R^2 describes the percentage of the variation of Y_i that is explained by the equation. The larger R^2, the larger this percentage

For this reason SSR is sometimes called the explained error while SSE is called the unexplained error.

Coefficient of correlation R:

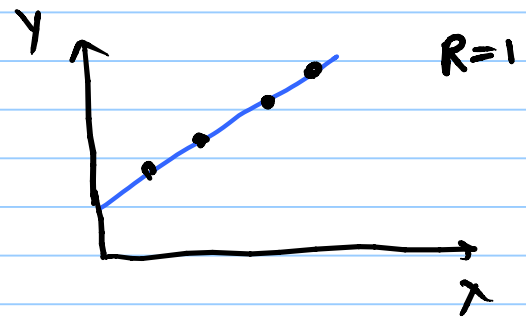
- The quantity

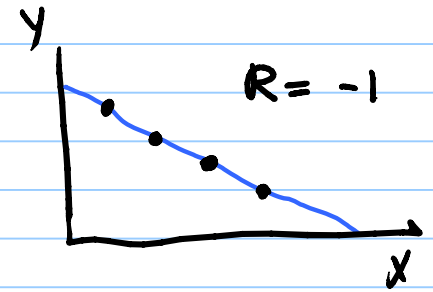
$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, SS_x SS_y} = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

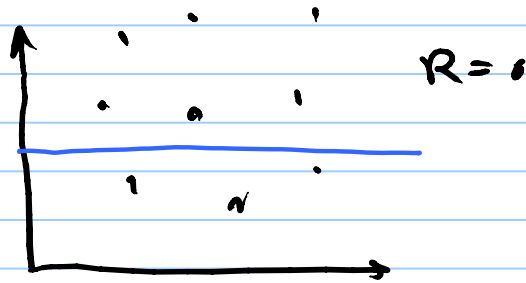
is called the **coefficient of correlation**

- It measure the **degree of linear relationship** between the x_i and the y_i: $-1 \leq R \leq 1$

If the point $(x_1, y_1), (x_2, y_2) \cdots, (x_n, y_n)$ are perfectly aligned then $R = 1$ or $R = -1$.

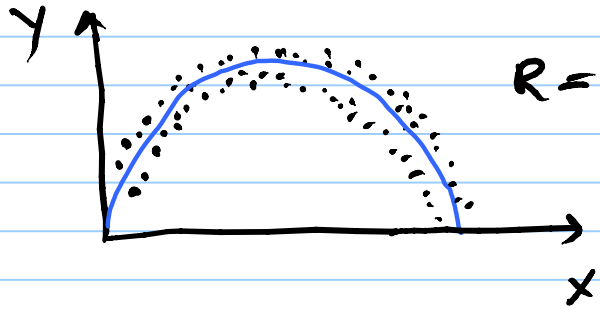






- If $R \approx 0$ it means there is little evidence of linear relationship between x and y.

  It does not mean there is no relationship!



- $Y = b_1 x(1-x) + b_0$ is a good fit
- $Y = b_0 + b_1 x$ is not

Confidence interval for estimating the response variable

- Once we have our estimated equation $\hat{y} = \hat{b}_0 + \hat{b}_1 x$ we can use it for new values of x.
- Point estimation is easy for a new x we get

$$\hat{y}_{new} = \hat{b}_0 + \hat{b}_1 x_{new}$$

- We are now interested in a confidence interval for our estimate
- There are two kinds of estimates.

1) Confidence interval for mean response $\bar{y}_{new}$:

Recall that our model is $y_{new} = b_0 + b_1 x_{new} + \epsilon$

- Thus if we make, say, k observations at x_{new} we'll get k possibly different y_{new} values.
- The mean of these $E(y_{new}) = b_0 + b_1 x_{new} = \bar{y}_{new}$
- It can be shown that the standard deviation is:

$$\sigma(\bar{Y}_{new}) = \sigma\sqrt{\frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{SS_x}}$$

But since σ is not known

$$SE(\bar{Y}_{new}) = S_e\sqrt{\frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{SS_x}}$$

} This quantity is the standard error of mean response

And the $1-\alpha$ confidence interval is

$$h = TINV(\alpha, n-2, 2)$$

Confidence interval $\hat{b}_0 + \hat{b}_1 X_{new} \pm h\ SE(\bar{Y}_{new})$

2) Confidence interval for predicted mean response

If we are not interested in the mean of response variable for the new value X_{new}, but rather the response variable itself then, still $E(Y_{pred}) = b_0 + b_1 X_{new}$

but

$$\sigma(Y_{pred}) = \sigma\sqrt{1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{SS_x}}$$

and when σ is not known

$$s(Y_{pred}) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{SS_x}}$$

} This quantity is standard error of prediction.

And confidence interval for predicted Y_{Pred} is:

$$\hat{b}_0 + \hat{b}_1 x_{new} \pm h\, s(Y_{Pred})$$

Some observations:

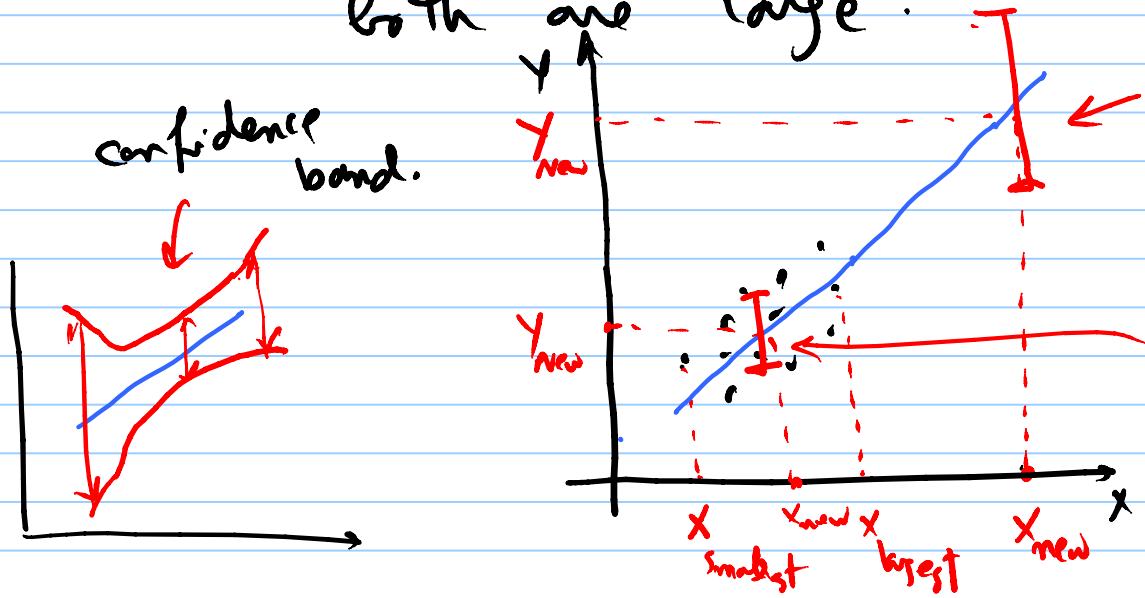
1) When X_{new} is near $\bar{X}$ the quantity $(X_{new} - \bar{x})^2$ is very small compared to $SS_x = \sum (x_i - \bar{x})^2$. So the standard error for $\bar{Y}_{new}$ is $s_E \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{SS_x}}$ is small. In fact, $SE(\bar{y}_{new}) \approx \sigma/\sqrt{n}$

But standard error for Y_{Pred}, $s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{SS_x}}$ is larger. This is natural because we can make a more accurate statement about the mean of

$SE(Y_{Pred}) \approx \sigma\sqrt{1+\frac{1}{n}} \approx \sigma$

population than an individual element of it.

2) When X_{new} is far from $\bar{X}$, in particular if X_{new} is much larger than the largest X_i or much smaller than the smallest X_i, then $X_{new} - \bar{X}$ is large and $S(E(Y_{new}))$ and $S(Y_{new})$ both are large.



Please study the examples in regressionSimple.xls file.