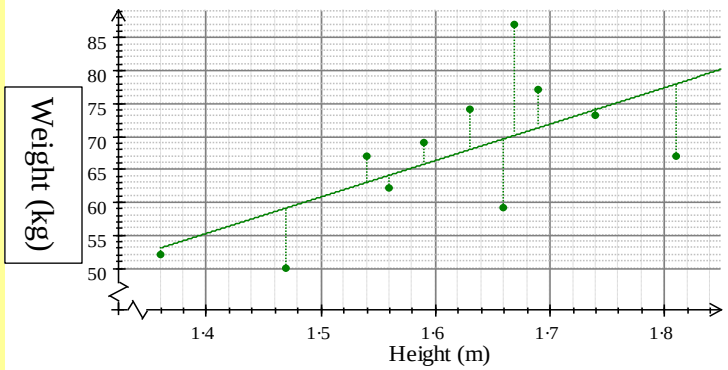
Simple Linear Regression:
1. Finding the equation of the line of best fit

Objectives: To find the equation of the least squares regression line of y on x.

Background and general principle

The aim of regression is to find the linear relationship between two variables. This is in turn translated into a mathematical problem of finding the equation of the line that is closest to all points observed.

Consider the **scatter plot** on the right. One possible line of best fit has been drawn on the diagram. Some of the points lie above the line and some lie below it.



The **vertical distance** each point is above or below the line has been added to the diagram. These distances are called *deviations* or *errors* – they are symbolised as $d_1, d_2, ..., d_n$.

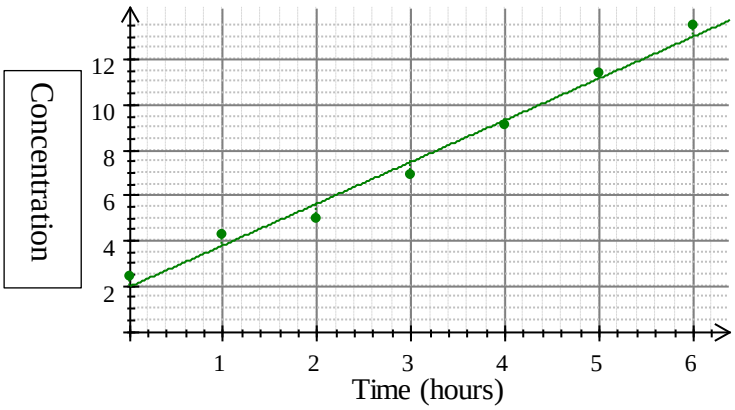
When drawing in a regression line, the aim is to make the line fit the points as closely as possible. We do this by making the **total of the squares of the deviations as small as possible**, i.e. we minimise $\sum d_i^2$.

If a line of best fit is found using this principle, it is called the **least-squares regression line**.

Example 1:
A patient is given a drip feed containing a particular chemical and its concentration in his blood is measured, in suitable units, at one hour intervals. The doctors believe that a linear relationship will exist between the variables.

Time, x (hours)	0	1	2	3	4	5	6
Concentration, y	2.4	4.3	5.0	6.9	9.1	11.4	13.5

We can plot these data on a scatter graph – time would be plotted on the horizontal axis (as it is the independent variable). Time is here referred to as a **controlled variable**, since the experimenter fixed the value of this variable in advance (measurements were taken every hour).



Concentration is the dependent variable as the concentration in the blood is likely to vary according to time.

The doctor may wish to estimate the concentration of the chemical in the blood after 3.5 hours.

She could do this by finding the equation of the line of best fit.

There is a formula which gives the equation of the line of best fit.

** The statistical equation of the simple linear regression line, when only the response variable Y is random, is: $Y = \beta_0 + \beta_1 x + \varepsilon$ (or in terms of each point: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$)
Here β_0 is called the intercept, β_1 the regression slope, ε is the random error with mean 0, x is the regressor (independent variable), and Y the response variable (dependent variable).

** The least squares regression line is obtained by finding the values of β_0 and β_1 values (denoted in the solutions as $\hat{\beta}_0$ & $\hat{\beta}_1$) that will minimize the sum of the squared vertical distances from all points to the line: $\Delta = \sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$
The solutions are found by solving the equations: $\frac{\partial \Delta}{\partial \beta_0} = 0$ and $\frac{\partial \Delta}{\partial \beta_1} = 0$

** The equation of the fitted least squares regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (or in terms of each point: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) ----- For simplicity of notations, many books denote the fitted regression equation as: $\hat{Y} = b_0 + b_1 x$ **(* you can see that for some examples, we will use this simpler notation.)**

where $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Notations: $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = \sum (x_i - \bar{x})(y_i - \bar{y})$; $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = \sum (x_i - \bar{x})^2$;
$\bar{x}$ and $\bar{y}$ are the mean values of x and y respectively.

Note 1: Please notice that **in finding the least squares regression line, we do not need to assume any distribution for the random errors ε_i. However, for statistical inference on the model parameters (** β_0 and β_1), it is assumed in our class that the errors have the following three properties:

- Normally distributed errors
- Homoscedasticity (constant error variance $\text{var}(\varepsilon_i) = \sigma^2$ for Y at all levels of X)
- Independent errors (usually checked when data collected over time or space)

***The above three properties can be summarized as: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2),\ i = 1, \cdots, n$

Note 2: Please notice that the least squares regression is only suitable when the random errors exist in the dependent variable Y only. If the regression X is also random – it is then referred to as the **Errors in Variable (EIV) regression**. One can find a good summary of the EIV regression in section 12.2 of the book: "Statistical Inference" (2nd edition) by George Casella and Roger Berger.

We can work out the equation for our example as follows:

$\sum x = 0 + 1 + \ldots + 6 = 21$ so $\bar{x} = \frac{21}{7} = 3$

$\sum y = 2.4 + 4.3 + \ldots + 13.5 = 52.6$ so $\bar{y} = \frac{52.6}{7} = 7.514\ldots$

$\sum xy = (0 \times 2.4) + (1 \times 4.3) + \ldots + (6 \times 13.5) = 209.4$

$\sum x^2 = 0^2 + 1^2 + \ldots + 6^2 = 91$ so $\bar{x} = \frac{21}{7} = 3$

These could all be found on a calculator (if you enter the data into a calculator).

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 209.4 - \frac{21 \times 52.6}{7} = 51.6$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 91 - \frac{(21)^2}{7} = 28$$

So, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{51.6}{28} = 1.843$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.514 - 1.843 \times 3 = 1.985$.

So the equation of the regression line is $\hat{y} = 1.985 + 1.843x$.

To work out the concentration after 3.5 hours: $\hat{y} = 1.985 + 1.843 \times 3.5 = 8.44$ (3sf)

If you want to find how long it would be before the concentration reaches 8 units, we substitute $\hat{y} = 8$ into the regression equation:

$8 = 1.985 + 1.843x$

Solving this we get: $x = 3.26$ hours

Note: It would not be sensible to predict the concentration after 8 hours from this equation – we don't know whether the relationship will continue to be linear. The process of trying to predict a value from outside the range of your data is called *extrapolation*.

Example 2:

The heights and weights of a sample of 11 students are:

Height (m) h	1.36	1.47	1.54	1.56	1.59	1.63	1.66	1.67	1.69	1.74	1.81
Weight (kg) w	52	50	67	62	69	74	59	87	77	73	67

$[n = 11 \quad \sum h = 17.72 \quad \sum h^2 = 28.705 \quad \sum w = 737 \quad \sum w^2 = 50571 \quad \sum hw = 1196.1]$

a) Calculate the regression line of w on h.
b) Use the regression line to estimate the weight of someone whose height is 1.6m.

Note: Both height and weight are referred to as **random** variables – their values could not have been predicted before the data were collected. If the sampling were repeated again, different values would be obtained for the heights and weights.

Solution:

a) We begin by finding the mean of each variable:

$$\bar{h} = \frac{\sum h}{n} = \frac{17.72}{11} = 1.6109...$$

$$\bar{w} = \frac{\sum w}{n} = \frac{737}{11} = 67$$

Next we find the sums of squares:

$$S_{hh} = \sum h^2 - \frac{\left(\sum h\right)^2}{n} = 28.705 - \frac{17.72^2}{11} = 0.1597$$

$$S_{ww} = \sum w^2 - \frac{\left(\sum w\right)^2}{n} = 50571 - \frac{737^2}{11} = 1192$$

$$S_{hw} = \sum hw - \frac{\sum h \sum w}{n} = 1196.1 - \frac{17.72 \times 737}{11} = 8.86$$

The equation of the least squares regression line is:

$$\hat{w} = \hat{\beta}_0 + \hat{\beta}_1 h$$

where

$$\hat{\beta}_1 = \frac{S_{hw}}{S_{hh}} = \frac{8.86}{0.1597} = 55.5$$

and

$$\hat{\beta}_0 = \bar{w} - \hat{\beta}_1 \bar{h} = 67 - 55.5 \times 1.6109 = -22.4$$

So the equation of the regression line of w on h is:

$$\hat{w} = -22.4 + 55.5h$$

b) To find the weight for someone that is 1.6m high:

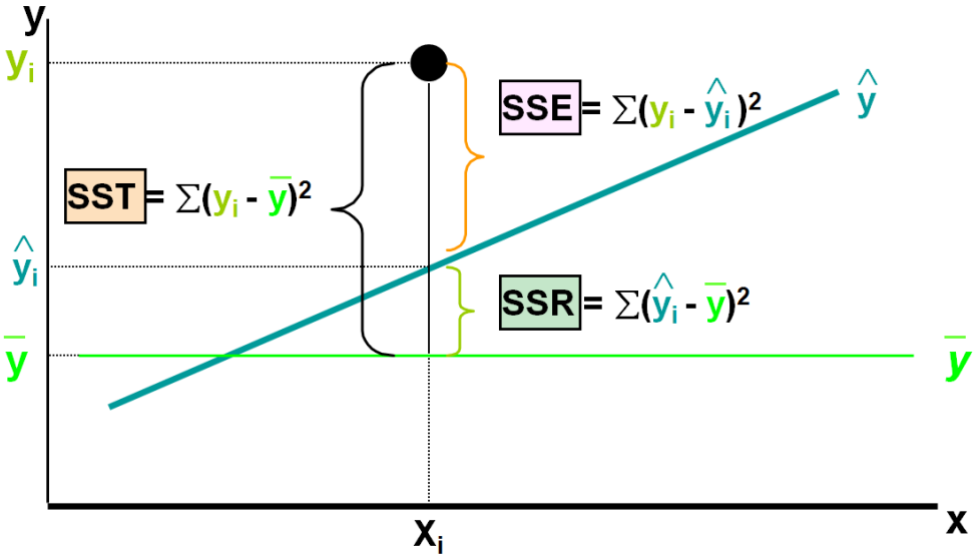
$$\hat{w} = -22.4 + 55.5 \times 1.6 = 66.4 \text{ kg}$$

Simple Linear Regression: 2. Measures of Variation

Objectives: measures of variation, the goodness-of-fit measure, and the correlation coefficient

Sums of Squares

- Total sum of squares = Regression sum of squares + Error sum of squares
  $SST = SSR + SSE$
  (Total variation = Explained variation + Unexplained variation)
- Total sum of squares (Total Variation): $SST = \sum\left(Y_i - \bar{Y}\right)^2$
- Regression sum of squares (Explained Variation by the Regression): $SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$
- Error sum of squares (Unexplained Variation): $SSE = \sum\left(Y_i - \hat{Y}_i\right)^2$



Coefficient of Determination and Correlation

Coefficient of Determination – it is a measure of the regression goodness-of-fit

It also represents the proportion of variation in Y "explained" by the regression on X

$$R^2 = \frac{SSR}{SST};\ 0 \le R^2 \le 1$$

Pearson (Product-Moment) Correlation Coefficient -- measure of the direction and strength of the linear association between *Y* and *X*

- For simple linear regression -- The sample correlation is denoted by ***r*** and is closely related to the coefficient of determination as follows:
- $r^2 = R^2$

$$r = sign\left(\hat{\beta}_1\right)\sqrt{R^2};\ -1 \le r \le 1$$

The sample correlation is indeed defined by the following formula:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2][\sum(y-\bar{y})^2]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{n\sum xy - \sum x\sum y}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}}$$

- The corresponding population correlation between Y and X is denoted by ρ and defined by:

$$\rho = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E\left[\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)\right]}{\sqrt{Var(X)Var(Y)}}$$

- Therefore one can see that in the population correlation definition, both X and Y are assumed to be random. When the joint distribution of X and Y is bivariate normal, one can perform the following t-test to test whether the population correlation is zero:

  - Hypotheses
    H_0: $\rho = 0$ (no correlation)
    H_A: $\rho \neq 0$ (correlation exists)

  - Test statistic

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \overset{H_0}{\sim} t_{n-2}$$

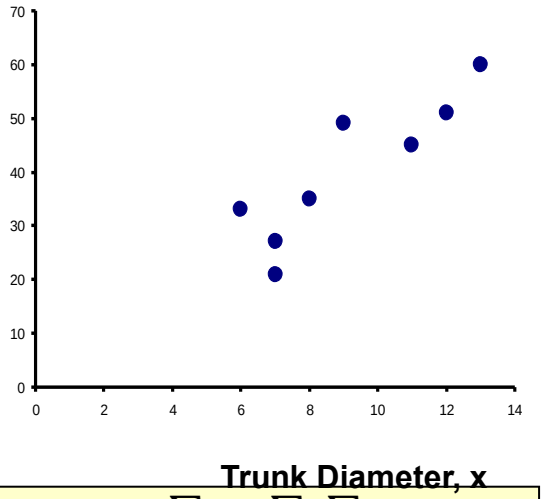
Note: One can show that this t-test is indeed the same t-test in testing the regression slope $\beta_1 = 0$ shown in the following section.

Note: The sample correlation is not an unbiased estimator of the population correlation. You can study this and other properties from the wiki site:
http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Example 3: The following example tabulates the relations between trunk diameter and tree height.

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	Σ=73	Σ=3142	Σ=14111	Σ=713

Scatter plot:



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Significance Test for Correlation

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

H_0: $\rho = 0$ (No correlation)
H_1: $\rho \neq 0$ (correlation exists)

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

At the significance level $\alpha = 0.05$, we reject the null hypothesis because $|t_0| = 4.68 \geq t_{6,0.025} = 2.447$ and conclude that there is a linear relationship between tree height and trunk diameter.

SAS for Correlation

```
Data tree;
Input height trunk;
Datalines;
35      8
49      9
27      7
33      6
60      13
21      7
45      11
51      12
;
Run;

Proc Corr data = tree;
Var height trunk;
Run;
```

Note: See the following website for more examples and interpretations of the output – plus how to draw the scatter plot (proc Gplot) in SAS: http://www.ats.ucla.edu/stat/sas/output/corr.htm

Standard Error of the Estimate (Residual Standard Deviation)

- The mean of the random error ε is equal to zero.
- An estimator of the standard deviation of the error ε is given by

$$\hat{\sigma} = s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Simple Linear Regression:
3. Inferences Concerning the Slope

Objectives: measures of variation, the goodness-of-fit measure, and the correlation coefficient

t-test

Test used to determine whether the population based slope parameter (β_1) is equal to a pre-determined value (often, but not necessarily 0). Tests can be one-sided (pre-determined direction) or two-sided (either direction).

2-sided t-test:

- H_0: $\beta_1 = 0$ (no linear relationship)
- H_1: $\beta_1 \neq 0$ (linear relationship does exist)

- **Test statistic:** $t_0 = \dfrac{b_1 - 0}{s_{b_1}}$ (Note: for simplicity of notations, $b_1 = \hat{\beta}_1$)

Where $s_{b_1} = \dfrac{s_\varepsilon}{\sqrt{S_{XX}}} = \dfrac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \dfrac{s_\varepsilon}{\sqrt{\sum x^2 - \dfrac{(\sum x)^2}{n}}}$

At the significance level α, we reject the null hypothesis if $|t_0| \geq t_{n-2,\alpha/2}$

(Note: one can also conduct the one-sided tests if necessary.)

F-test (based on *k* independent variables)

A test based directly on sum of squares that tests the specific hypotheses of whether the slope parameter is 0 (2-sided). The book describes the general case of *k* predictor variables, **for simple linear regression, *k* = 1**.

$$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$$

$$TS: F_{obs} = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

$$RR: F_{obs} \geq F_{\alpha,k,n-k-1}$$

Analysis of Variance (based on *k* Predictor Variables – for simple linear regression, *k* = 1)

Source	**df**	**Sum of Squares**	**Mean Square**	**F**
Regression	*k*	*SSR*	*MSR=SSR/k*	$F_{obs}=MSR/MSE$
Error	*n-k*-1	*SSE*	*MSE=SSE/(n-k*-1)	---
Total	*n*-1	*SST*	---	---

100(1-α)% Confidence Interval for the slope parameter, β_1: $b_1 \pm t_{n-2,\alpha/2} s_{b_1}$

(Note: for simplicity of notations, $b_1 = \hat{\beta}_1$)

- If entire interval is positive, conclude $\beta_1 > 0$ (Positive association)
- If interval contains 0, conclude (do not reject) $\beta_1 = 0$ (No association)
- If entire interval is negative, conclude $\beta_1 < 0$ (Negative association)

Example 4: A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet). A random sample of 10 houses is selected

- Dependent variable (y) = house price in $1000s
- Independent variable (x) = square feet
-

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Solution: Regression analysis output:

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Estimated house price = 98.24833 + 0.10977 (square feet)

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X (Note: for simplicity of notations, $b_1 = \hat{\beta}_1$)
  - Here, b_1 = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

This means that 58.08% of the variation in house prices is explained by variation in square feet.

$$s_\varepsilon = 41.33032$$

The standard error (estimated standard deviation of the random error) is also given in the output (above).

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic:
  - $$t_0 = \frac{b_1 - 0}{s_{b_1}} \approx 3.329$$ (Note: for simplicity of notations, $b_1 = \hat{\beta}_1$)

At the significance level $\alpha = 0.05$, we reject the null hypothesis because $|t_0| = 3.329 \geq t_{8,0.025} = 2.306$ and conclude that there is sufficient evidence that square footage affects house price.

Confidence Interval Estimate of the Slope: $b_1 \pm t_{n-2,\alpha/2} s_{b_1}$
(Note: for simplicity of notations, $b_1 = \hat{\beta}_1$)
The 95% confidence interval for the slope is (0.0337, 0.1858)

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.
Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Predict the price for a house with 2000 square feet:

$$\begin{aligned} \text{house price} &= 98.25 + 0.1098\,(\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85 \end{aligned}$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,85

Example 5 (SAS): What is the relationship between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	Birthweight (g/1000)
1	1
2	1
3	2
4	2
5	4

```
Data BW;  /*Reading data in SAS*/
input estriol birthw@@;
datalines;
1 1 2 1 3 2 4 2 5 4
;
run;

PROC REG data=BW; /*Fitting linear regression models*/
model birthw=estriol;
run;
```

Finance Application: Market Model

- One of the most important applications of linear regression is the *market model*.
- It is assumed that rate of return on a stock (R) is linearly related to the rate of return on the overall market.

$$R = \beta_0 + \beta_1 R_m + \varepsilon$$

R: Rate of return on a particular stock
R_m: Rate of return on some major stock index
β_1: The beta coefficient measures how sensitive the stock's rate of return is to changes in the level of the overall market.

Example: Here we estimate the market model for Nortel, a stock traded in the Toronto Stock Exchange. Data consisted of monthly percentage return for Nortel and monthly percentage return for all the stocks.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.560079
R Square	0.313688
Adjusted F	0.301855
Standard E	0.063123
Observatio	60

ANOVA

	df	*SS*	*MS*	*F*	*ignificance*
Regressior	1	0.10563	0.10563	26.50969	3.27E-06
Residual	58	0.231105	0.003985		
Total	59	0.336734			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	0.012818	0.008223	1.558903	0.12446
TSE	0.887691	0.172409	5.148756	3.27E-06

TSE (estimated regression slope): This is a measure of the stock's market related risk. In this sample, for each 1% increase in the TSE return, the average increase in Nortel's return is .8877%.

R Square (R^2) This is a measure of the total risk embedded in the Nortel stock, that is market-related. Specifically, 31.37% of the variation in Nortel's return are explained by the variation in the TSE's returns.

Linear Regression in Matrix Form

Data:

$$(y_1, x_{11}, x_{12}, \ldots, x_{1p-1}), (y_2, x_{21}, x_{22}, \ldots, x_{2p-1}), \ldots, (y_n, x_{n1}, x_{n2}, \ldots, x_{np-1}).$$

The **multiple linear regression model** in scalar form is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.$$

The above linear regression can also be represented in the vector/matrix form. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

Then,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1p-1} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_{p-1} x_{2p-1} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{np-1} + \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1p-1} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_{p-1} x_{2p-1} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{np-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} =$$

$$\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Estimation:

Least square method:

The least square method is to find the estimate of $\boldsymbol{\beta}$ minimizing the sum of square of residual,

$$S(\boldsymbol{\beta}) = S(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \varepsilon_i^2 = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon}$$

$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

since $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Expanding $S(\boldsymbol{\beta})$ yields

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}^t - \boldsymbol{\beta}^t \mathbf{X}^t)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$= \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

$$= \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

Note:

For two matrices A and B, $(AB)^t = B^t A^t$ and $\left(A^{-1}\right)^t = \left(A^t\right)^{-1}$

Similar to the procedure in finding the minimum of a function in calculus, the least square estimate ***b*** can be found by solving the equation based on the first derivative of $S(\boldsymbol{\beta})$,

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \partial S(\boldsymbol{\beta})/\partial \beta_0 \\ \partial S(\boldsymbol{\beta})/\partial \beta_1 \\ \vdots \\ \partial S(\boldsymbol{\beta})/\partial \beta_{p-1} \end{bmatrix} = \frac{\partial(\mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\Leftrightarrow \quad \mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}$$

$$\Leftrightarrow \mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$ (Note, here we denote **b** as the estimator of **β**, that is $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}$)

The fitted regression equation is

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}$. (Note: for simplicity, here: $b_i = \hat{\beta}_i, i = 0, .., p-1$)

The fitted values (in vector): $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$

The residuals (in vector): $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

Note: (i) $\frac{\partial(\boldsymbol{\beta}^t\mathbf{a})}{\partial \boldsymbol{\beta}} = \frac{\partial(\sum_{i=1}^{p} \beta_{i-1} a_i)}{\partial \boldsymbol{\beta}} = \mathbf{a}$, where $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ and $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$.

(ii) $\frac{\partial(\boldsymbol{\beta}^t\mathbf{A}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial(\sum_{i=1}^{p}\sum_{j=1}^{p} \beta_{i-1}\beta_{j-1}a_{ij})}{\partial \boldsymbol{\beta}} = 2\mathbf{A}\boldsymbol{\beta}$, where A is any symmetric $p \times p$ matrix.

Note: Since

$$(X^t X)^t = X^t (X^t)^t = X^t X,$$

$X^t X$ is a symmetric matrix.

Also,

$$((X^t X)^{-1})^t = ((X^t X)^t)^{-1} = (X^t X)^{-1},$$

$(X^t X)^{-1}$ is a symmetric matrix.

Note: $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$ ***is called the normal equation.***

Note: $\mathbf{e}^t\mathbf{X} = (\mathbf{y}^t - \mathbf{b}^t\mathbf{X}^t)\mathbf{X} = [\mathbf{y}^t - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]\mathbf{X} = \mathbf{y}^t\mathbf{X} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}$

$= \mathbf{y}^t\mathbf{X} - \mathbf{y}^t\mathbf{X} = \mathbf{0}$.

Therefore, if there is intercept, then the first column of X is $\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$. Then,

$$\mathbf{e}^t\mathbf{X}=\begin{bmatrix}e_1 & e_2 & \cdots & e_n\end{bmatrix}\begin{bmatrix}1 & x_{11} & \cdots & x_{1p-1}\\1 & x_{21} & \cdots & x_{2p-1}\\\vdots & \vdots & \ddots & \vdots\\1 & x_{n1} & \cdots & x_{np-1}\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}e_i & \cdots\end{bmatrix}=\mathbf{0}$$

$$\Rightarrow\sum_{i=1}^{n}e_i=0$$

Note: for the linear regression model without the intercept, $\sum_{i=1}^{n}e_i$ might not be equal to 0.

Properties of the least square estimate:

Two useful results:

Let $Z_{n\times1}=\begin{bmatrix}Z_1\\Z_2\\\vdots\\Z_n\end{bmatrix}$ **be a** $n\times1$ **random vector,** $A_{p\times n}$ **is a** $p\times n$ **matrix and** $C_{n\times1}$ **is a** $n\times1$ **vector. Let**

$$E(Z)=\begin{bmatrix}E(Z_1)\\E(Z_2)\\\vdots\\E(Z_n)\end{bmatrix}$$

and

$$V(Z)=\begin{bmatrix}\mathrm{cov}(Z_1,Z_1) & \mathrm{cov}(Z_1,Z_2) & \cdots & \mathrm{cov}(Z_1,Z_n)\\\mathrm{cov}(Z_2,Z_1) & \mathrm{cov}(Z_2,Z_2) & \cdots & \mathrm{cov}(Z_2,Z_n)\\\vdots & \vdots & \ddots & \vdots\\\mathrm{cov}(Z_n,Z_1) & \mathrm{cov}(Z_n,Z_2) & \cdots & \mathrm{cov}(Z_n,Z_n)\end{bmatrix}$$

$$=\begin{bmatrix}Var(Z_1) & \mathrm{cov}(Z_1,Z_2) & \cdots & \mathrm{cov}(Z_1,Z_n)\\\mathrm{cov}(Z_2,Z_1) & Var(Z_2) & \cdots & \mathrm{cov}(Z_2,Z_n)\\\vdots & \vdots & \ddots & \vdots\\\mathrm{cov}(Z_n,Z_1) & \mathrm{cov}(Z_n,Z_2) & \cdots & Var(Z_n)\end{bmatrix}.$$

Then

(a) $E(AZ)=AE(Z),\ E(Z+C)=E(Z)+C$.

(b) $V(AZ)=AV(Z)A^t,\ V(Z+C)=V(Z)$

Note:

$$E(\boldsymbol{\varepsilon})=\mathbf{0},V(\varepsilon)=\sigma^2I$$

The properties of least square estimate:

1. $E(\mathbf{b}) = \begin{bmatrix} E(b_0) \\ E(b_1) \\ \vdots \\ E(b_{p-1}) \end{bmatrix} = \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$

2. The variance –covariance matrix of the least square estimate b is

$$V(\mathbf{b}) = \begin{bmatrix} Var(b_0) & \text{cov}(b_0, b_1) & \cdots & \text{cov}(b_0, b_{p-1}) \\ \text{cov}(b_1, b_0) & Var(b_1) & \cdots & \text{cov}(b_1, b_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_{p-1}, b_0) & \text{cov}(b_{p-1}, b_1) & \cdots & Var(b_{p-1}) \end{bmatrix}$$
$$= \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}$$

[Derivation:]

$$E(\mathbf{b}) = E\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\right] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E(\mathbf{y}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

since

$$E(\mathbf{y}) = E\left[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right] = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}.$$

Also,

$$V(\mathbf{b}) = V\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\right] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t V(\mathbf{y}) \left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t$$
$$= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}$$
$$= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

since

$$V(\mathbf{y}) = V\left[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right] = V(\boldsymbol{\varepsilon}) = \sigma^2 I$$

Example 6: (This is a simple example of multiple regression.)

Heller Company manufactures lawn mowers and related lawn equipment. The managers believe the quantity of lawn mowers sold depends on the price of the mower and the price of a competitor's mower. We have the following data:

Competitor's Price x_{i1}	*Heller's Price* x_{i2}	*Quantity sold* y_i
120	**100**	**102**
140	**110**	**100**
190	**90**	**120**
130	**150**	**77**
155	**210**	**46**
175	**150**	**93**
125	**250**	**26**
145	**270**	**69**
180	**300**	**65**
150	**250**	**85**

The ***regression model*** for the above data is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i .$$

The data in matrix form are

$$\mathbf{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix}=\begin{bmatrix} 102 \\ 100 \\ \vdots \\ 85 \end{bmatrix},\mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{101} & x_{102} \end{bmatrix}=\begin{bmatrix} 1 & 120 & 100 \\ 1 & 140 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 150 & 250 \end{bmatrix},\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

The least square estimate $\mathbf{b}$ is

$$\mathbf{b}=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}=\begin{bmatrix} 66.518 \\ 0.414 \\ -0.269 \end{bmatrix}.$$

The fitted regression equation is

$$\hat{y}=b_0+b_1x_1+b_2x_2=66.518+0.414x_1-0.269x_2.$$

The fitted equation implies ***an increase*** in the competitor's price of 1 unit is associated with ***an increase*** of 0.414 unit in expected quantity sold and ***an increase*** in its own price of 1 unit is associated with ***a decrease*** of 0.269 unit in expected quantity sold.

Suppose now we want to predict the quantity sold in a city where Heller prices it mower at \$160 and the competitor prices its mower at \$170. The quantity sold predicted is

$$66.518+0.414\cdot 170-0.269\cdot 160=93.718.$$

Example 7:

We show how to use the matrix approach to obtain the least square estimate and its expected value and the variance.

Let

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i, i=1,\ldots,n.$$

Then,

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1\,x_1 \\ 1\,x_2 \\ \vdots\ \vdots \\ 1\,x_n \end{bmatrix}, \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Thus,

$$X^tX=\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} 1\,x_1 \\ \vdots\ \vdots \\ 1\,x_n \end{bmatrix}=\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix},\ X^ty=\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{bmatrix}$$

and

$$(X^tX)^{-1}=\frac{1}{n\sum_{i=1}^n x_i^2-(\sum_{i=1}^n x_i)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}=\frac{1}{\sum_{i=1}^n x_i^2-n\bar{x}^2}\begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$=\frac{1}{S_{XX}}\begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix},$$

since

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Therefore,

$$b = (X^t X)^{-1} X^t y = \frac{1}{S_{XX}}\begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \frac{1}{S_{XX}}\begin{bmatrix} \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i}{n} - \left(\sum_{i=1}^n x_i y_i\right)\bar{x} \\ -\bar{x}\sum_{i=1}^n y_i + \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$= \frac{1}{S_{XX}}\begin{bmatrix} \bar{y}\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) + n\bar{x}^2\bar{y} - \left(\sum_{i=1}^n x_i y_i\right)\bar{x} \\ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{bmatrix}$$

$$= \frac{1}{S_{XX}}\begin{bmatrix} \bar{y}S_{XX} - \bar{x}\left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right) \\ S_{XY} \end{bmatrix} = \frac{1}{S_{XX}}\begin{bmatrix} \bar{y}S_{XX} - \bar{x}S_{XY} \\ S_{XY} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x}\frac{S_{XY}}{S_{XX}} \\ \frac{S_{XY}}{S_{XX}} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} - b_1\bar{x} \\ b_1 \end{bmatrix}$$

Also,

$$E(b) = \begin{bmatrix} E(b_0) \\ E(b_1) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

and

$$V(b) = \begin{bmatrix} Var(b_0) & \mathrm{cov}(b_0, b_1) \\ \mathrm{cov}(b_0, b_1) & Var(b_1) \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{nS_{XX}} & -\frac{\bar{x}}{S_{XX}} \\ -\frac{\bar{x}}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix}\sigma^2$$

Acknowledgement: In compiling this lecture notes, we have revised some materials from the following websites:
www.schoolworkout.co.uk/documents/s1/Regression.doc
www.stat.ufl.edu/~winner/mar5621/mar5621.doc
www.fordham.edu/economics/Vinod/correl-regr.ppt
www2.thu.edu.tw/~wenwei/Courses/regression/ch6.1.doc
www.msu.edu/~fuw/teaching/Fu_Ch11_linear_regression.ppt
www.stanford.edu/class/msande247s/kchap17.ppt