Linear Regression

Fifth session

Simple Linear Regression

Learning Objectives

- Introduce the straight-line (simple linear regression) model as a means of relating one quantitative variable to another quantitative variable
- Introduce the correlation coefficient as a means of relating one quantitative variable to another quantitative variable
- Assess how well the simple linear regression model fits the sample data
- Employ the simple linear regression model for predicting the value of one variable from a specified value of another variable

An example

Age and cholesterol levels in 18 individuals

ID	Age	Chol (mg/ml)
1	46	3.5
2	20	1.9
3	52	4.0
4	30	2.6
5	57	4.5
6	25	3.0
7	28	2.9
8	36	3.8
9	22	2.1
10	43	3.8
11	57	4.1
12	33	3.0
13	22	2.5
14	63	4.6
15	40	3.2
16	48	4.2
17	28	2.3
18	49	4.0

Questions of interest

- Association between age and cholesterol levels
- Strength of association
- Prediction of cholesterol for a given age

Correlation and Regression analysis

Variance and covariance: algebra

- Let *x* and *y* be two random variables from a sample of *n* observations.
- Measure of variability of *x* and *y*: **variance**

$$\mathrm{var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad \mathrm{var}(y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

- Measure of covariation between *x* and *y* ?

- Algebraically:

  var(*x* ± *y*) = var(*x*) + var(*y*)

  var(*x* ± *y*) = var(*x*) + var(*y*) ± 2cov(*x*,*y*)

Where:

$$\mathrm{cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Meaning of variance and covariance

- Variance is always positive
- If covariance = 0, *x* and *y* are independent.
- Covariance is sum of cross-products: can be positive or negative.
- Negative covariance = deviations in the two distributions in are opposite directions, e.g. genetic covariation.
- Positive covariance = deviations in the two distributions in are in the same direction.
- Covariance = a measure of strength of association.

Covariance and correlation

- Covariance is unit-depenent.
- Coefficient of correlation (*r*) between *x* and *y* is a standardized covariance.
- *r* is defined by:

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x) \times \operatorname{var}(y)}} = \frac{\operatorname{cov}(x, y)}{SD_x \times SD_y}$$

Coefficient of Correlation

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

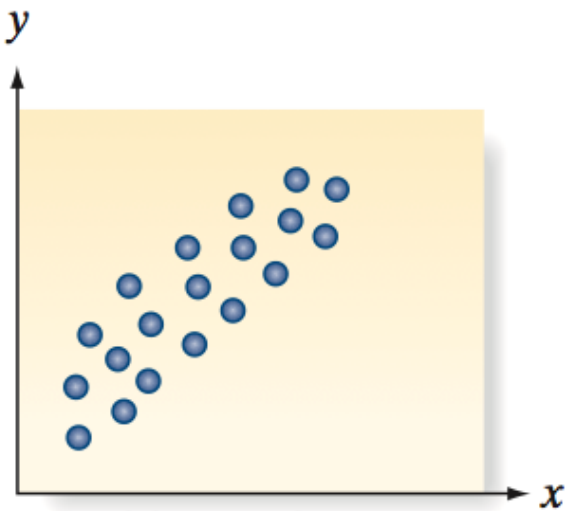
where

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

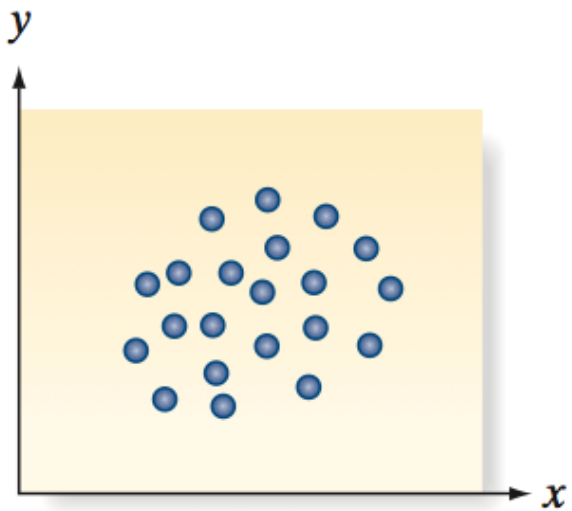
$$SS_{xx} = \sum (x - \bar{x})^2$$

$$SS_{yy} = \sum (y - \bar{y})^2$$

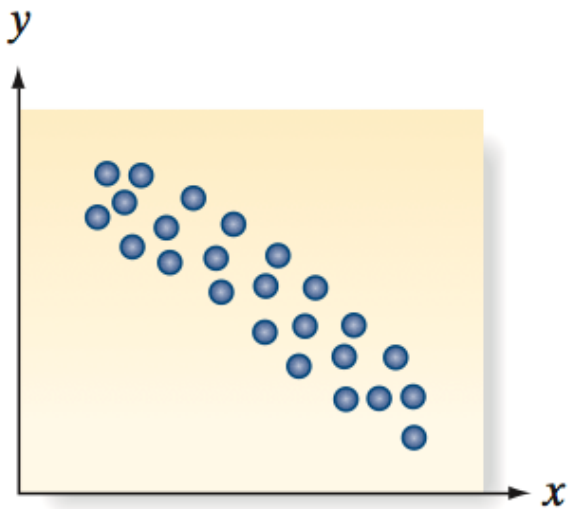
Coefficient of Correlation



a. Positive r: y increases as x increases

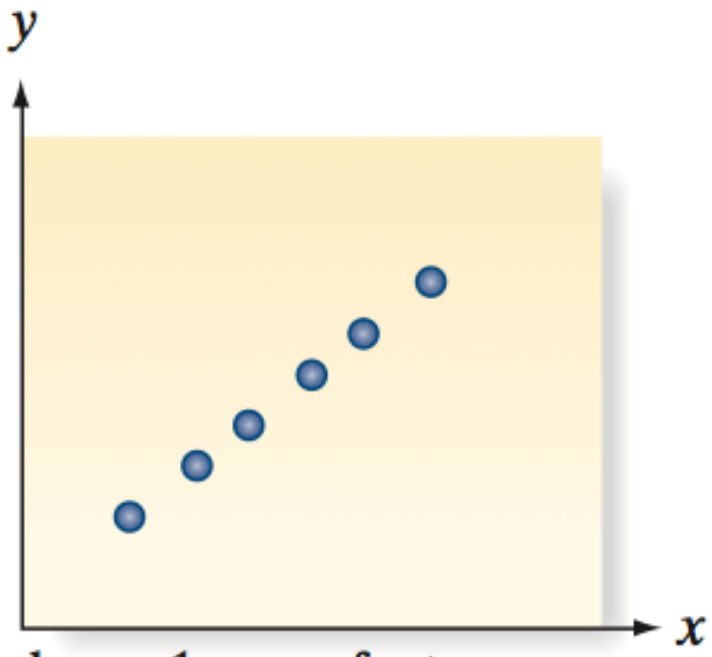


b. r near 0: little or no relationship between y and x

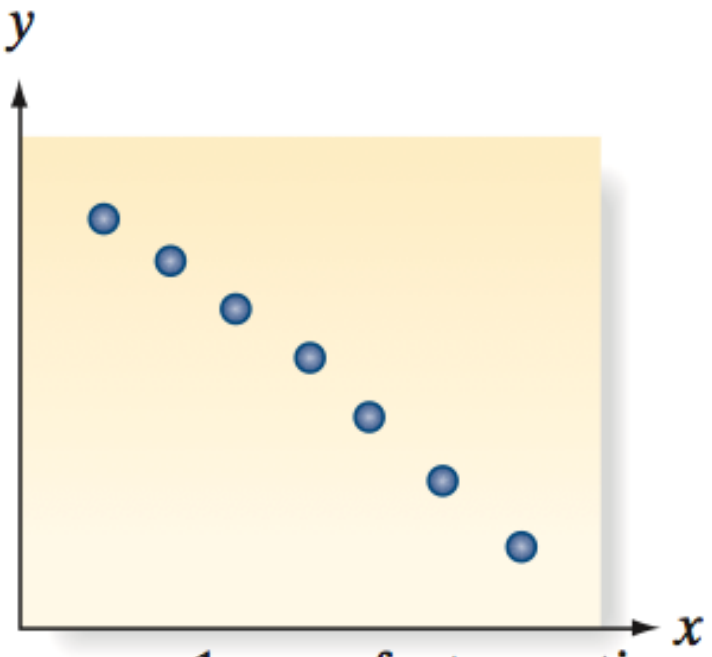


c. Negative r: y decreases as x increases

Coefficient of Correlation

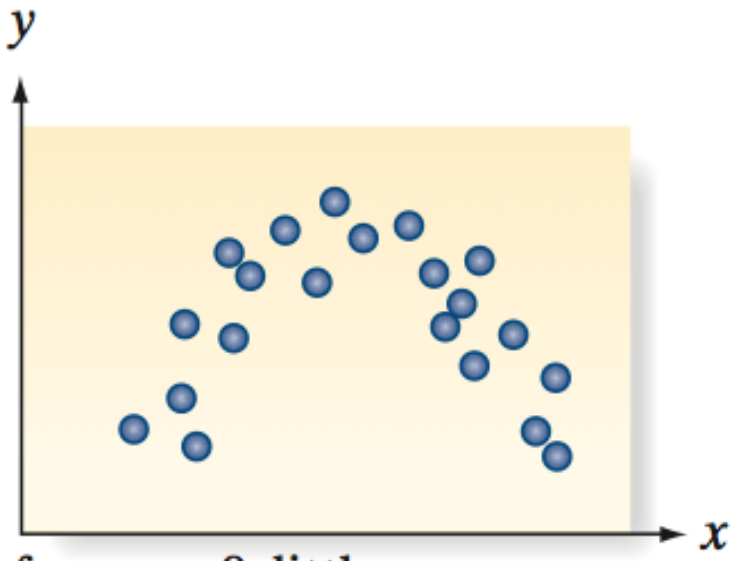


d. $r = 1$: a perfect positive relationship between y and x



e. $r = -1$: a perfect negative relationship between y and x

Coefficient of Correlation



f. r near 0: little or no linear relationship between y and x

Test of hypothesis of correlation

- Hypothesis: H_o: $r = 0$ versus H_o: r not equal to 0.

- Standard error of r is: $$SE(r) = \sqrt{\frac{1-r^2}{n-2}}$$

- The t-statistic: $$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- This statistic has a t distribution with $n - 2$ degrees of freedom.

Test of hypothesis of correlation

- Hypothesis: H_o: r = r0 versus H_o: r ≠ r0.

- Fisher's z-transformation: $$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$
- Standard error of z: $$SE(z) = \frac{1}{\sqrt{n-3}}$$
- Then 95% CI of z can be constructed as: $$z \pm \frac{1}{\sqrt{n-3}}$$

An illustration of correlation analysis

ID	Age (x)	Cholesterol (y; mg/100ml)
1	46	3.5
2	20	1.9
3	52	4.0
4	30	2.6
5	57	4.5
6	25	3.0
7	28	2.9
8	36	3.8
9	22	2.1
10	43	3.8
11	57	4.1
12	33	3.0
13	22	2.5
14	63	4.6
15	40	3.2
16	48	4.2
17	28	2.3
18	49	4.0
Mean	**38.83**	**3.33**
SD	**13.60**	**0.84**

$$\mathrm{Cov}(x, y) = 10.68$$

$$r = \frac{\mathrm{cov}(x,y)}{SD_x \times SD_y} = \frac{10.68}{13.60 \times 0.84} = 0.94$$

$$z = \frac{1}{2}\ln\left(\frac{1+0.94}{1-0.94}\right) = 0.56$$

$$SE(z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{15}} = 0.26$$

t-statistic = 0.56 / 0.26 = 2.17

Critical t-value with 17 df and alpha = 5% is 2.11

Conclusion: There is a significant association between age and cholesterol.

Simple linear regression analysis

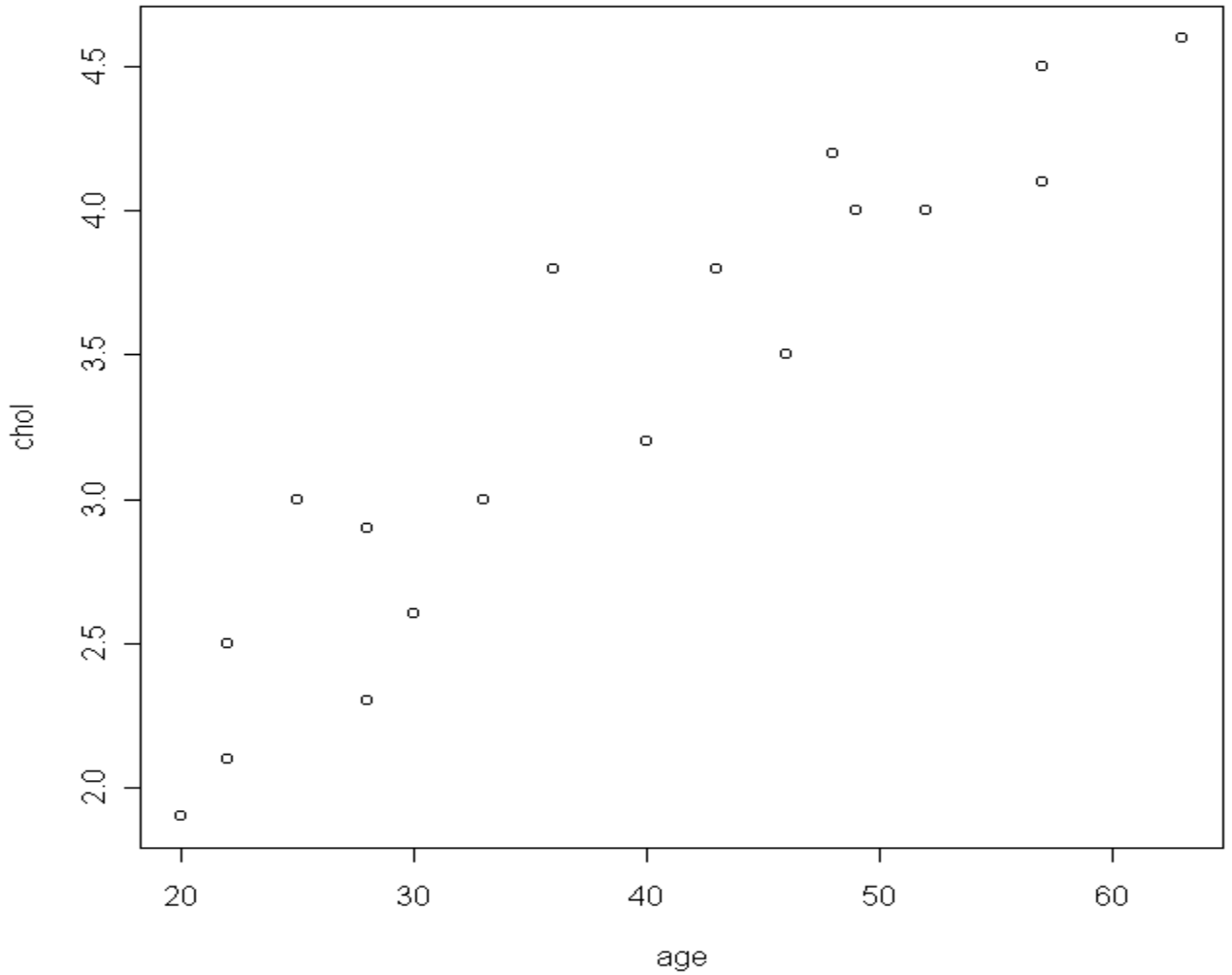
In general, simple linear regression finds the best straight line for describing the relationship between two variables. In its simplest form, which is what we consider here, it does not do a very good job of assessing how well the line describes the data, but nevertheless provides useful information.

Simple linear regression analysis

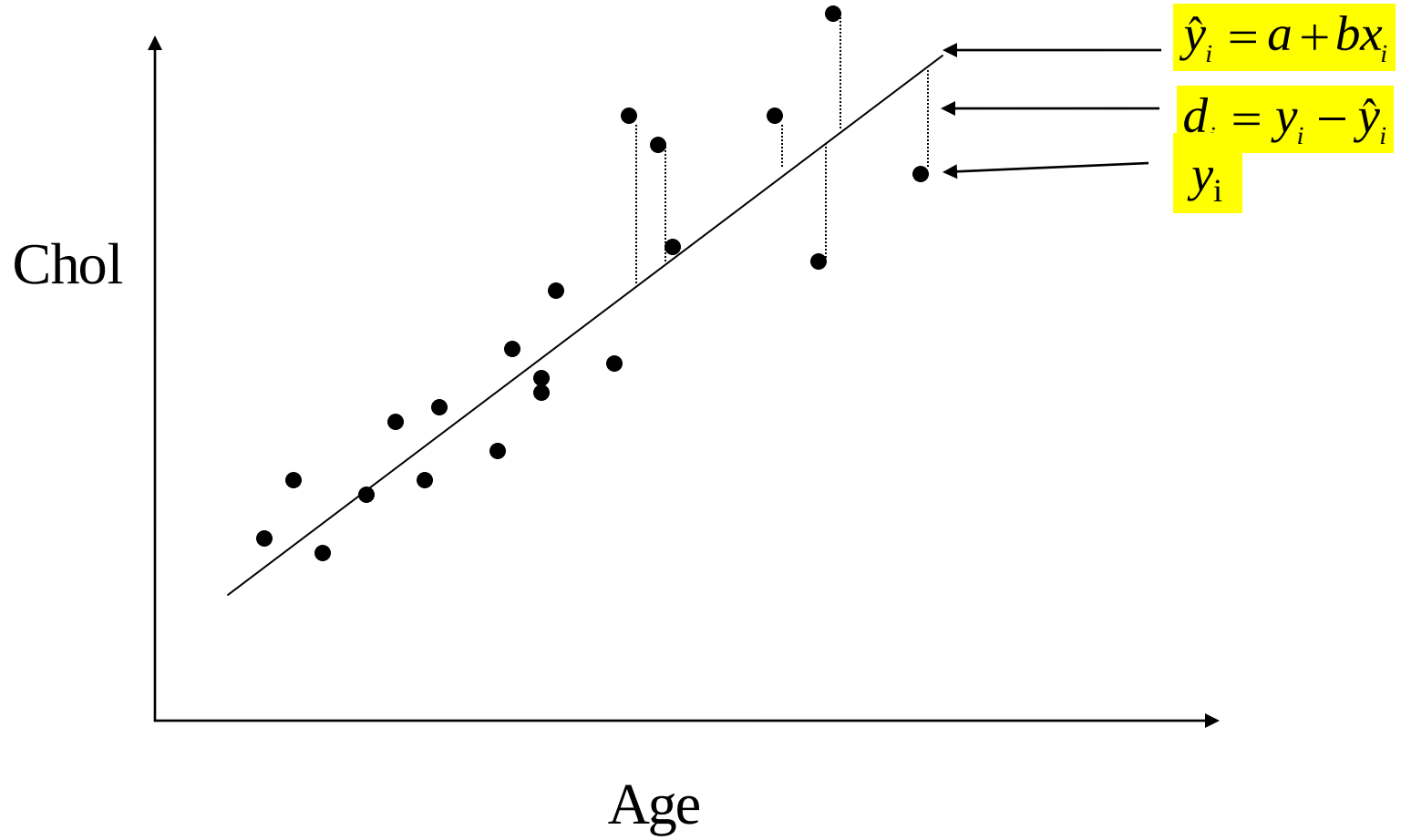
- Only two variables are of interest: one response variable and one predictor variable
- No adjustment is needed for confounding or covariate

- Assessment:
  - Quantify the relationship between two variables
- Prediction
  - Make prediction and validate a test
- Control
  - Adjusting for confounding effect (in the case of multiple variables)

Relationship between age and cholesterol



Criteria of estimation



The goal of least square estimator (LSE) is to find a and b such that the sum of d^2 is minimal.

Linear regression: model

- *Y* : random variable representing a **response**
- *X* : random variable representing a **predictor** variable (predictor, risk factor)
  - *Y is a* continuous variable (e.g., chol level) and *X* can be a categorical variable (e.g., yes / no) or a continuous variable (e.g., age).

- Model

$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

β : slope / gradient

ε : random error (variation between subjects in y even if *x* is constant, e.g., variation in cholesterol for patients of the same age.)

Linear regression: assumptions

- The relationship is linear *in terms of the parameter*;
- *X* is measured without error;
- The values of *Y* are independently from each other (e.g., Y_1 is not correlated with Y_2) ;
- The random error term (e) is **normally** distributed with mean 0 and **constant** variance.

Expected value and variance

- If the assumptions are tenable, then:
- The expected value of Y is: $E(Y \mid x) = \alpha + \beta x$
- The variance of Y is: $\text{var}(Y) = \text{var}(\varepsilon) = \sigma^2$

Estimation of a and b

- For a series of pairs: (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n)
- Let a and b be **sample estimates** for **parameters** α and β,
- We have a sample equation: $Y^* = a + bx$

- Aim: finding the values of a and b so that $(Y - Y^*)$ is minimal.

- Let SSE = sum of $(Y_i - a - bx_i)^2$.
- Values of a and b that minimise SSE are called **least square estimates.**

Estimation of a and b

- After some calculus operations, the results can be shown to be:

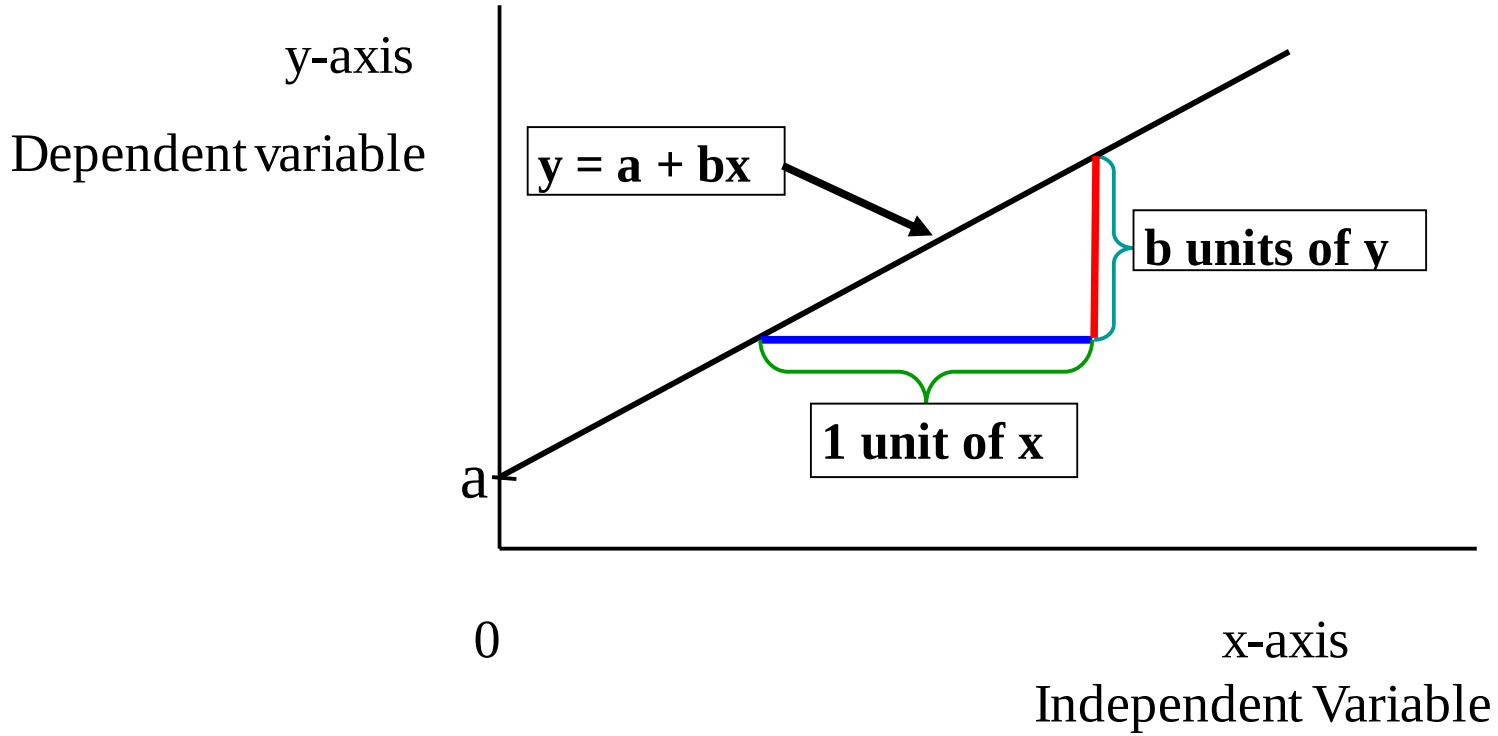
$$a = \bar{y} - b\bar{x}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

Where: $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- When the regression assumptions are valid, the estimators of α and β have the following properties:
  - Unbiased
  - Uniformly minimal variance (eg efficient)



a = Intercept, that is, the point where the line crosses the y-axis, which is the value of y at **x** = 0.

b = Slope of the regression line, that is, the number of units of increase (positive slope) or decrease (negative slope) in **y** for each unit increase in **x**.

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of b_1 is zero.

- Two tests are commonly used:

  ***t* Test** and ***F* Test**

- Both the *t* test and *F* test require an estimate of s^2, the variance of *e* in the regression model.

Regression ANOVA

If the regression line is flat in the sense that the regression estimate of Y, being $\hat{y}$, is the same for all values of x, then there is no gain from considering the x variable as it is having no impact on $\hat{y}$. This situation occurs when the estimated slope b = 0. An important question is whether or not the population parameter $\beta = 0$, that is, whether the truth is that there is no linear relationship between y and x. To test this situation, we can proceed with a formal test.

1. The Hypothesis: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. The α level: $\alpha = 0.05$

3. The assumptions: Random normal samples for y-variable from populations defined by x-variable

4. The test statistic:

ANOVA				
Source	df	SS	MS	F
Regression	1	SS(*Reg*)	SS(*Reg*)/1	MS(*Reg*)/MS(*Res*)
Residual	n-2	SS(*Res*)	SS(*Res*)/(n-2)	
Total	n-1	SS(y)		

5. The rejection region : Reject $H_0: \beta = 0$ if the value calculated for F is greater than $F_{0.95}(1, n-2)$

Partitioning of variations: concept

- SST = sum of squared difference between y_i and the mean of y.
- SSR = sum of squared difference between the predicted value of y and the mean of y.
- SSE = sum of squared difference between the observed and predicted value of y.

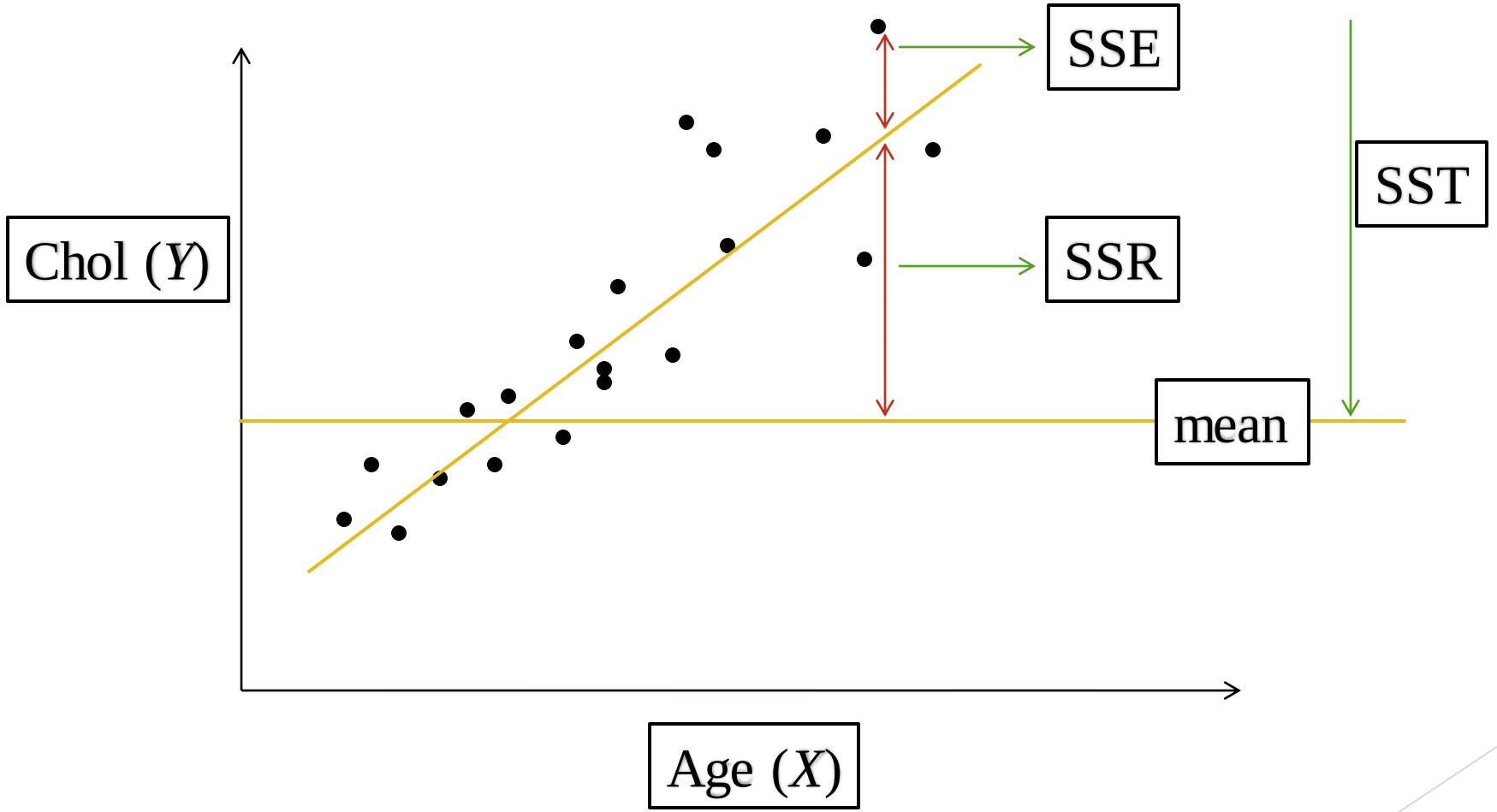
SST = SSR + SSE

The the coefficient of determination is: R^2 = SSR / SST

$$r = \hat{\beta}_1 \left(\frac{S_{xx}}{SS_T} \right)^{1/2}$$

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

Partitioning of variations: geometry



Partitioning of variations: algebra

- Some statistics:
  - Total variation: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - Attributed to the model: $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - Residual sum of square: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
  - $SST = SSR + SSE$
  - $SSR = SST - SSE$

t tests in regression analysis

- Now, we have

  Sample data: $Y = a + bX + e$

  Population: $Y = a + bX + e$

- H_o: b = 0. There is no linear association between the outcome and predictor variable.

- In layman language: “what is the chance, given the sample data that we observed, of observing a sample of data that is less consistent with the null hypothesis of no association?”

Inference about slope (parameter β)

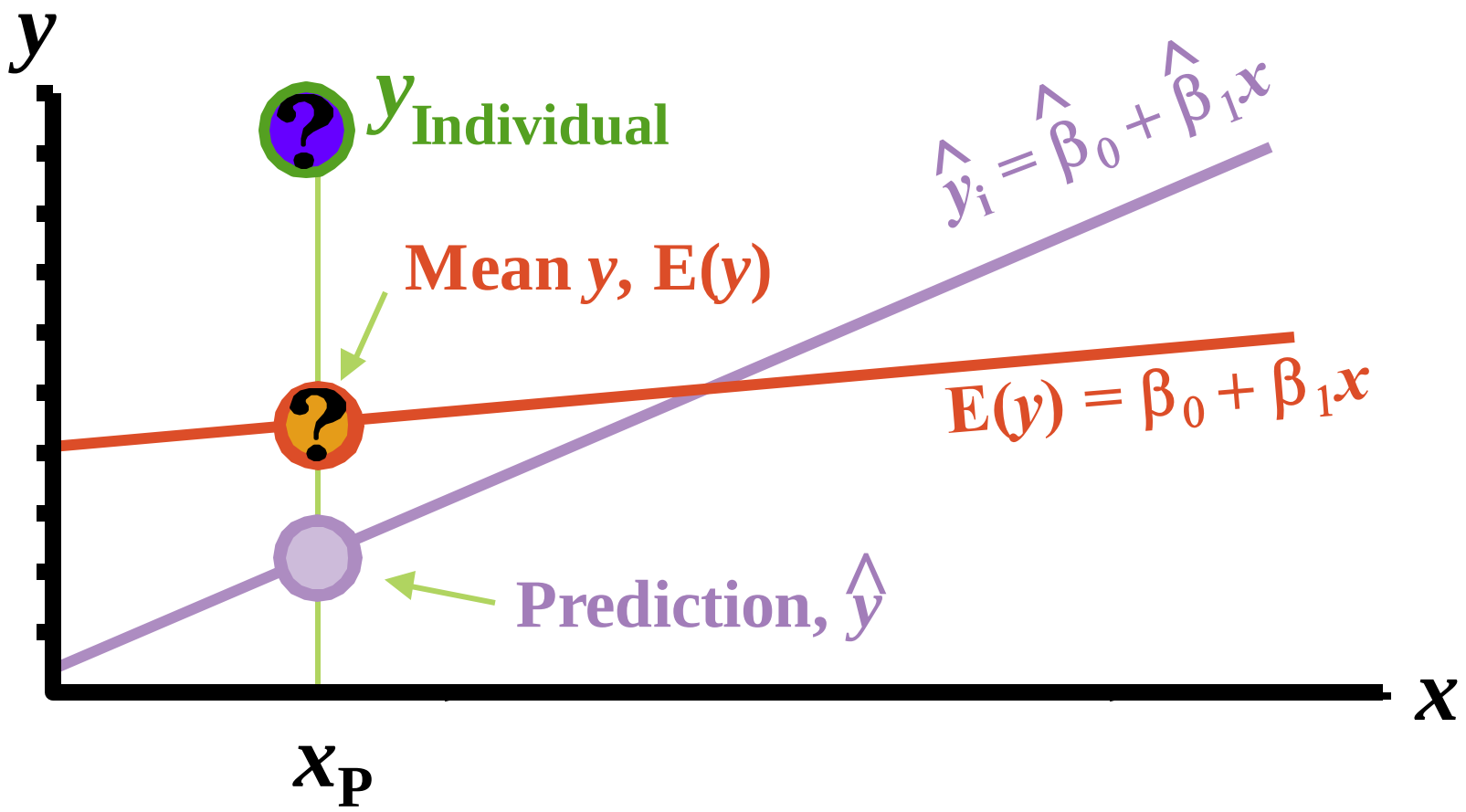
- Recall that e is assumed to be normally distributed with mean 0 and variance = s^2.
- Estimate of s^2 is MSE (or s^2)
- It can be shown that
  - The expected value of *b* is b, i.e. E(*b*) = b,
  - The standard error of *b* is: $SE(b) = s / \sqrt{S_{xx}}$

- Then the test whether b = 0 is: $t = b / SE(b)$ which follows a t-distribution with n-1 degrees of freedom.

Prediction With Regression Models

- Types of predictions
  - Point estimates
  - Interval estimates
- What is predicted
  - Population mean response $E(y)$ for given x
    - Point on population regression line
  - Individual response (y_i) for given x

What Is Predicted



A 100(1 - α)% Confidence Interval Estimate for the Mean Value of y at $x = x_p$

$$\ddot{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

$$df = n - 2$$

Confidence interval around predicted valued

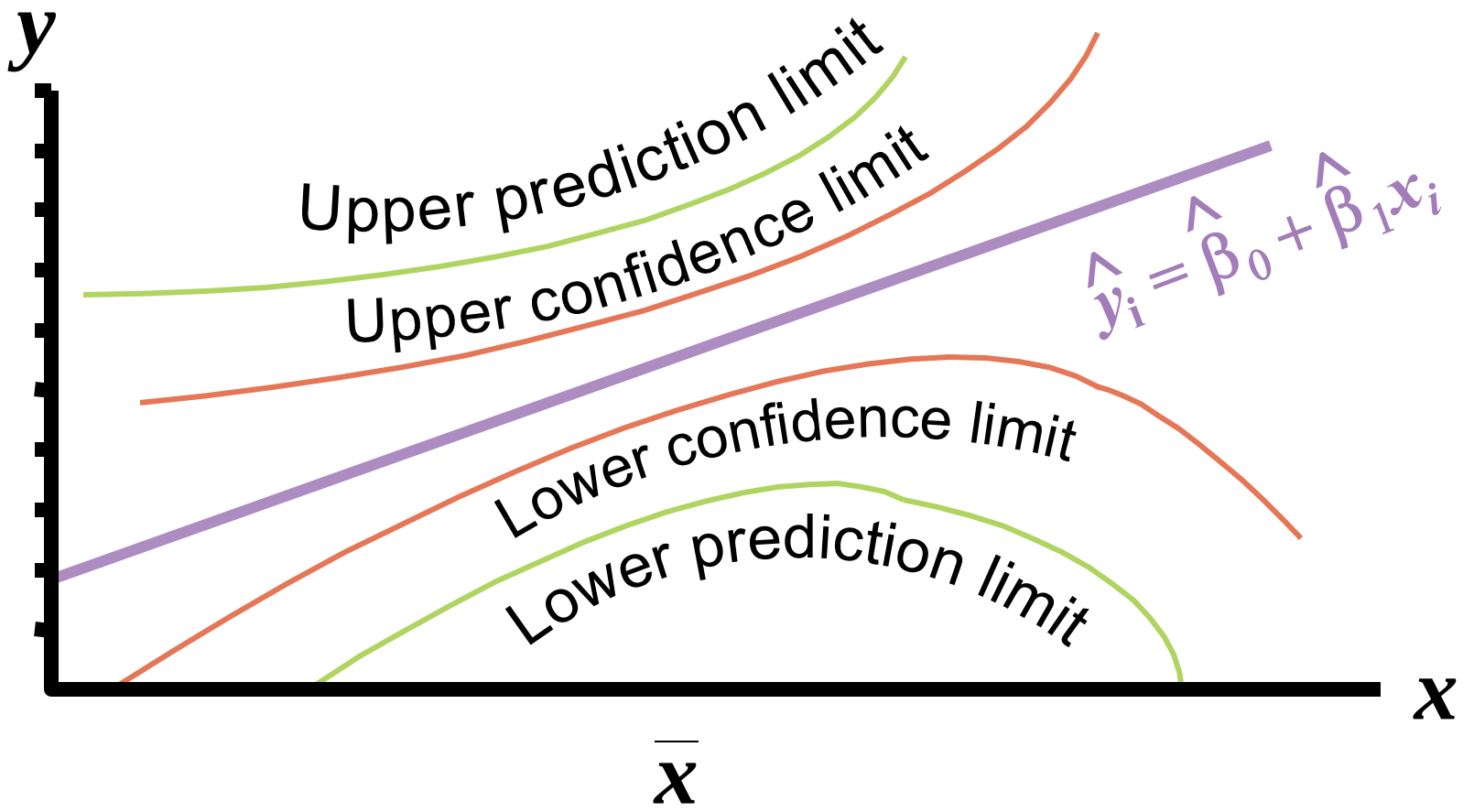
- Observed value is Y_i.
- Predicted value is $\hat{Y}_i = a + bx_i$
- The standard error of the predicted value is:

$$SE\left(\hat{Y}_i\right) = s\sqrt{1 + \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{S_{xx}}}$$

- Interval estimation for Y_i values

$$\hat{Y}_i \pm SE\left(\hat{Y}_i\right) \times \left(t_{n-p-1,1-\alpha/2}\right)$$

Confidence Intervals v. Prediction Intervals

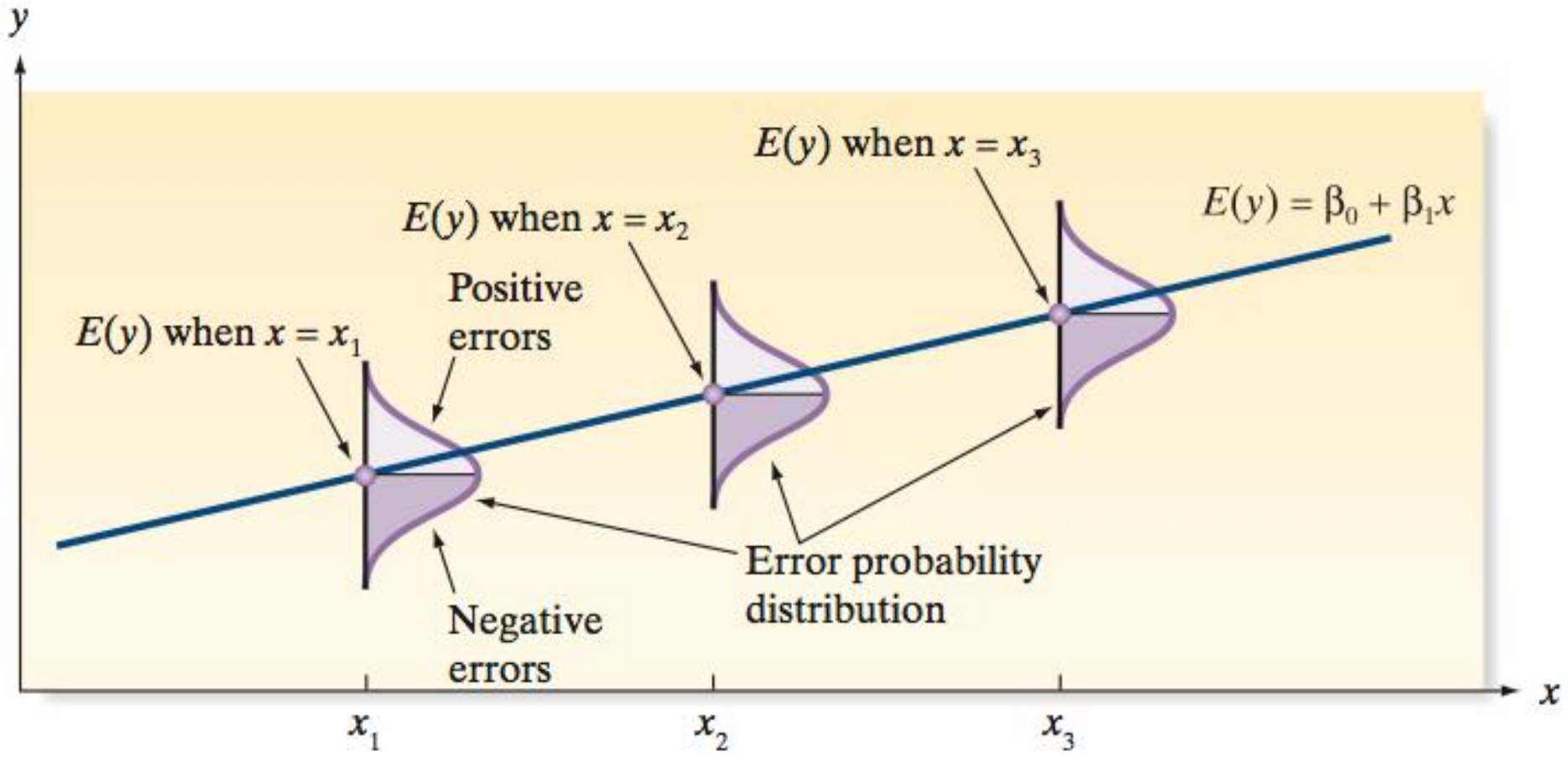


Checking assumptions

- Assumption of constant variance
- Assumption of normality
- Correctness of functional form
- Model stability

- All can be conducted with graphical analysis. The residuals from the model or a function of the residuals play an important role in all of the model diagnostic procedures.

Basic Assumptions of the Probability Distribution



Checking assumptions

- Assumption of constant variance
  - Plot the studentized residuals versus their predicted values. Examine whether the variability between residuals remains relatively constant across the range of fitted values.
- Assumption of normality
  - Plot the residuals versus their expected values under normality (Normal probability plot). If the residuals are normally distributed, it should fall along a 45^{o} line.
- Correct functional form?
  - Plot the residuals versus fitted values. Examine whether the residual plot for evidence of a non-linear trend in the value of the residual across the range of fitted values.
- Model stability
  - Check whether one or more observations are influential. Use Cook's distance.

Checking assumptions (Cont)

- **Cook's distance** (D) is a measure of the magnitude by which the fitted values of the regression model change if the ith observation is removed from the data set.

- **Leverage** is a measure of how extreme the value of x_i is relative to the remaining value of x.

- The **Studentized residual** provides a measure of how extreme the value of y_i is relative to the remaining value of y.

Some comments: Interpretation of correlation

- Correlation lies between –1 and +1. A very small correlation does *not* mean that no linear association between the two variables. The relationship may be non-linear.
- For curlinearity, a rank correlation is better than the Pearson's correlation.
- A small correlation (eg 0.1) may be statistically significant, but clinically unimportant.
- R^2 is another measure of strength of association. An $r = 0.7$ may sound impressive, but R^2 is 0.49!
- Correlation does not mean causation.

Example:

The following data are diastolic blood pressure (DBP) measurements taken at different times after an intervention for n = 5 persons. For each person, the data available include the time of the measurement and the DBP level. Of interest is the relationship between these two variables.

	Time		DPB		
Patient	x	x^2	y	y^2	xy
1	0	0	72	5,184	0
2	5	25	66	4,356	330
3	10	100	70	4,900	700
4	15	225	64	4,096	960
5	20	400	66	4,356	1,320
Sum	50	750	338	22,892	3,310
Mean	10		67.6		
n	5		5		

For the blood pressure data,

$$\bar{x} = 50/5 = 10,$$

$$\bar{y} = 338/5 = 67.6,$$

the slope is

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

$$b = \frac{3,310 - (50)(338)/5}{750 - (50)^2/5} = -0.28$$

and the intercept is

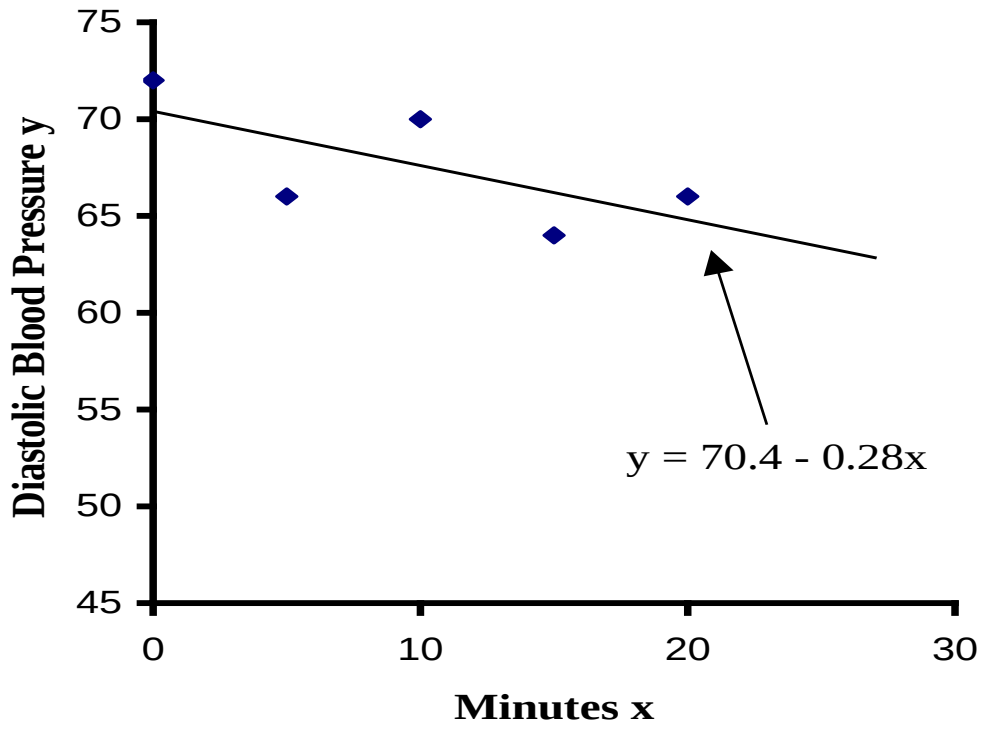
$$a = \bar{y} - b\bar{x},$$

$$a = 67.6 - (-0.28)10 = 70.4$$

The best line is

$$y = a + bx = 70.4 - 0.28x$$

Patient	Time x	DBP y
1	0	72
2	5	66
3	10	70
4	15	64
5	20	66



Blood Pressure Example

$$SS(Total) = SS(y) = \sum (y - \bar{y})^2$$

$$= 22{,}892 - \frac{(338)^2}{5} = 43.2$$

$$SS(Regression) = bSS(xy)$$

$$= b\left\{\sum xy - \sum x \sum y / n\right\}$$

$$= -0.28\{3310 - (50)(338)/5\} = 19.6$$

$$SS(Residual) = SS(Total) - SS(Regression)$$

$$= 43.2 - 19.6 = 23.6$$

ANOVA

Source	df	SS	MS	F
Regression	1	19.6	19.6	2.49
Residual	3	23.6	7.89	
Total	4	43.2		

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

For $\alpha = 0.05$ $F_{0.95(1,3)} = 10.1$, Hence accept $H_0 : \beta = 0$

$$R^2 = \frac{SS(Regression)}{SS(Total)} = \frac{19.6}{43.2} = 0.4537 \quad \text{or} \quad 45.37\%$$

Note: The above hypothesis test does not asses how well the straight line fits the data.

Multiple linear regression

Multiple Linear regression

Multiple Linear Regression Model

In a **multiple linear regression model,** the **dependent variable** or **response** is related to k **independent** or **regressor variables.** The model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (6\text{-}3)$$

Estimation of Parameters in Multiple Regression

The method of least squares may be used to estimate the regression coefficents in the multiple regression model, equation 6-3. Suppose that $n > k$ observations are available, and let x_{ij} denote the ith observation or level of variable x_j. The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) \quad i = 1, 2, \ldots, n > k$$

It is customary to present the data for multiple regression in a table such as Table 6-4.

Table 6-4 Data for Multiple Linear Regression

y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Estimation of Parameters in Multiple Regression

- The least squares normal equations are

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

the solution to the normal equations are the least squares estimators of the regression coefficients.

Multiple Regression example:

Table 6-5 Wire Bond Pull Strenght Data for Example 6-7

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

Multiple Regression example:

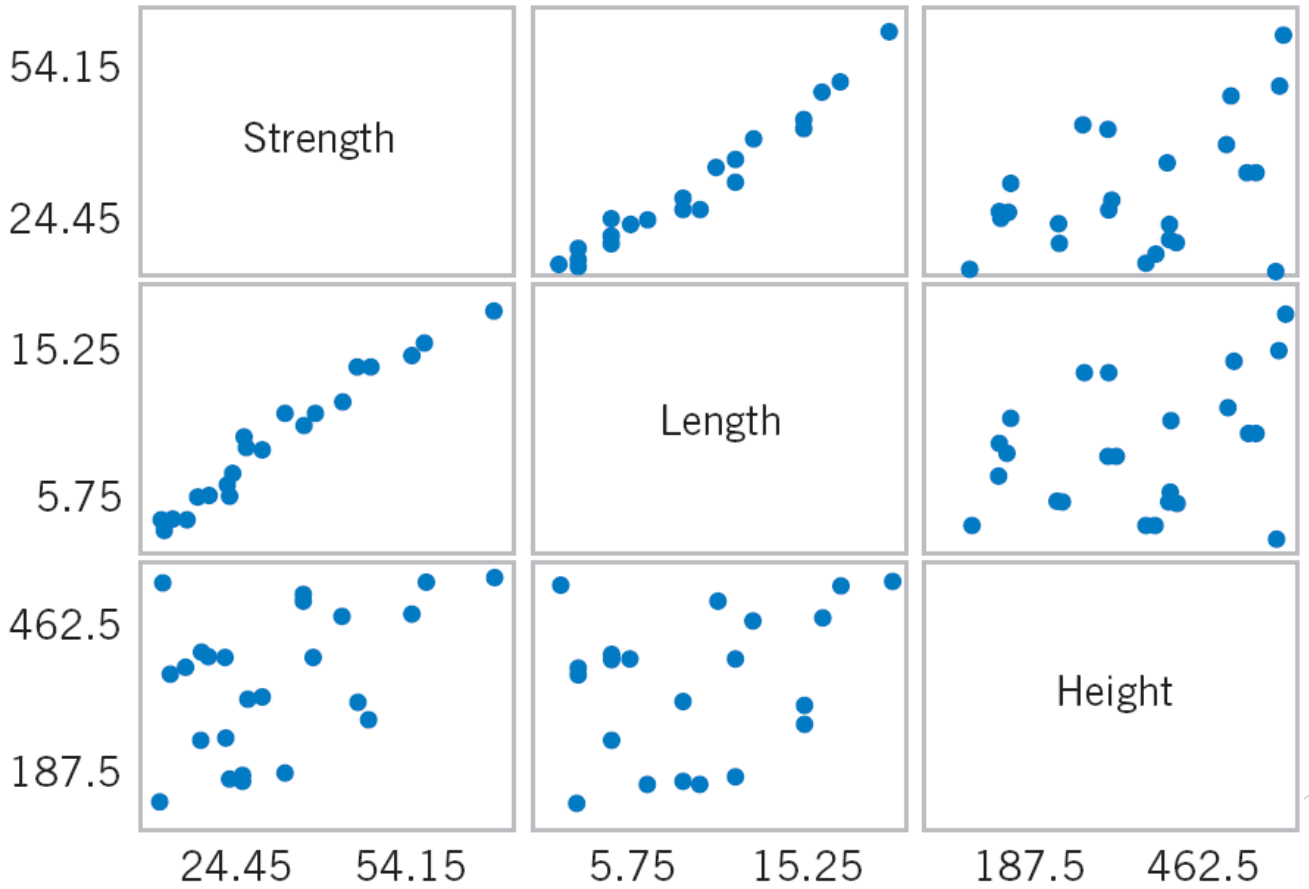


Figure 6-17 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 6-5.

Estimation of Parameters in Multiple Regression

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} = \frac{SS_E}{n - p} \qquad (6\text{-}45)$$

The Adjusted Coefficient of Multiple Determination (R^2_{Adjusted})

The **adjusted coefficient of multiple determination** for a multiple regression model with k regressors is

$$R^2_{\text{Adjusted}} = 1 - \frac{SS_E/(n - p)}{SS_T/(n - 1)} = \frac{(n - 1)R^2 - k}{n - p} \qquad (6\text{-}46)$$

Analysis of variance

- SS increases in proportion to sample size (n)
- Mean squares (MS): normalise for degrees of freedom (df)
  - $MSR = SSR / p$ (where p = number of degrees of freedom)
  - $MSE = SSE / (n - p - 1)$
  - $MST = SST / (n - 1)$
- Analysis of variance (ANOVA) table:

Source	d.f.	Sum of squares (SS)	Mean squares (MS)	F-test
Regression	p	SSR	MSR	MSR/MSE
Residual	N–p –1	SSE	MSE	
Total	n – 1	SST		

Inferences in Multiple Regression

Test for Significance of Regression

Testing for Significance of Regression in Multiple Regression

$$MS_R = \frac{SS_R}{k} \qquad MS_E = \frac{SS_E}{n-p}$$

Null hypothesis: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Alternative hypothesis: H_1: At least one $\beta_j \neq 0$

Test statistic: $F_0 = \dfrac{MS_R}{MS_E}$ (6-47)

P-value: Probability above f_0 in the $F_{k,n-p}$ distribution

Rejection criterion for a fixed-level test: $f_0 > f_{\alpha,k,n-p}$

Inferences in Multiple Regression

Inference on Individual Regression Coefficients

Inferences on the Model Parameters in Multiple Regression

1. The test for $H_0: \beta_j = \beta_{j,0}$ versus $H_1: \beta_j \neq \beta_{j,0}$ employs the **test statistic**

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{se(\hat{\beta}_j)} \qquad (6\text{-}48)$$

and the null hypothesis is rejected if $|t_0| > t_{\alpha/2,n-p}$. A P-value approach can also be used. One-sided alternative hypotheses can also be tested.

2. A $100(1-\alpha)\%$ CI for an individual regression coefficient is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p}se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}se(\hat{\beta}_j) \qquad (6\text{-}49)$$

- This is called a **partial** or marginal test

Inferences in Multiple Regression

Confidence Intervals on the Mean Response and Prediction Intervals

Confidence Interval on the Mean Response in Multiple Regression

A $100(1-\alpha)\%$ CI on the mean response at the point $x_1 = x_{10}, x_2 = x_{20}, \ldots, x_k = x_{k0}$ in a multiple regression model is given by

$$\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}} - t_{\alpha/2,n-p}\,se(\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}}) \le \mu_{Y|x_{10},x_{20},\ldots,x_{k0}} \le \hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}} + t_{\alpha/2,n-p}\,se(\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}}) \quad (6\text{-}52)$$

where $\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}}$ is computed from equation 6-51.

Inferences in Multiple Regression

Confidence Intervals on the Mean Response and Prediction Intervals

Prediction Interval on a Future Observation

A $100(1-\alpha)\%$ PI on a future observation at the point $x_1 = x_{10}, x_2 = x_{20}, \ldots, x_k = x_{k0}$ in a multiple regression model is given by

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 + [se(\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}})]^2} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 + [se(\hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}})]^2} \quad (6\text{-}54)$$

where $\hat{y}_0 = \hat{\mu}_{Y|x_{10},x_{20},\ldots,x_{k0}}$ is computed from equation 6-53.

Inferences in Multiple Regression

A Test for the Significance of a Group of Regressors

$$H_0: \beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$$

$$H_1: \text{At least one of the } \beta\text{'s} \neq 0$$

we would use the test statistic

$$F_0 = \frac{[SS_E(RM) - SS_E(FM)]/(k - r)}{SS_E(FM)/(n - p)}$$