



İST 292 STATISTICS

Sections: 05-06

For Department of Computer Engineering

LESSON 8 REGRESSION ANALYSIS


Dr. Ayten Yiğiter and Dr. Esra Polat Lecture Notes

REGRESSION ANALYSIS

Many engineering and scientific problems are concerned with determining a relationship between a set of variables.

Regression analysis, is a statistical technique that is very useful for these types of problems.

For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis, can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes.



Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. In regression analysis, the dependent variable is denoted "Y" and the independent variables are denoted by "X". The variable that is being predicted (the variable that the equation solves for) is called the dependent variable. The variable/variables that are used to predict the value of the dependent variable are called the independent variables.

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable.



Examples:

- Monthly natural gas consumption (Y) and average temperature (X).
- Family income (X) and expence amount (Y)
- Yield (Y) and amount of precipitation (yağış miktarı) (X).
- Final grade (Y) and average study hours weekly (X).
- Price of cigarette (X) and amount of sales (Y)

Table 1. Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level (x)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

As an illustration, consider data in Table 1. In this table **y** is the purity of oxygen produced in a chemical distillation (damıtma) process, and **x** is the percentage of hydrocarbons that are present in the main condenser (biriktirici, yoğunlaştırıcı) of the distillation unit.

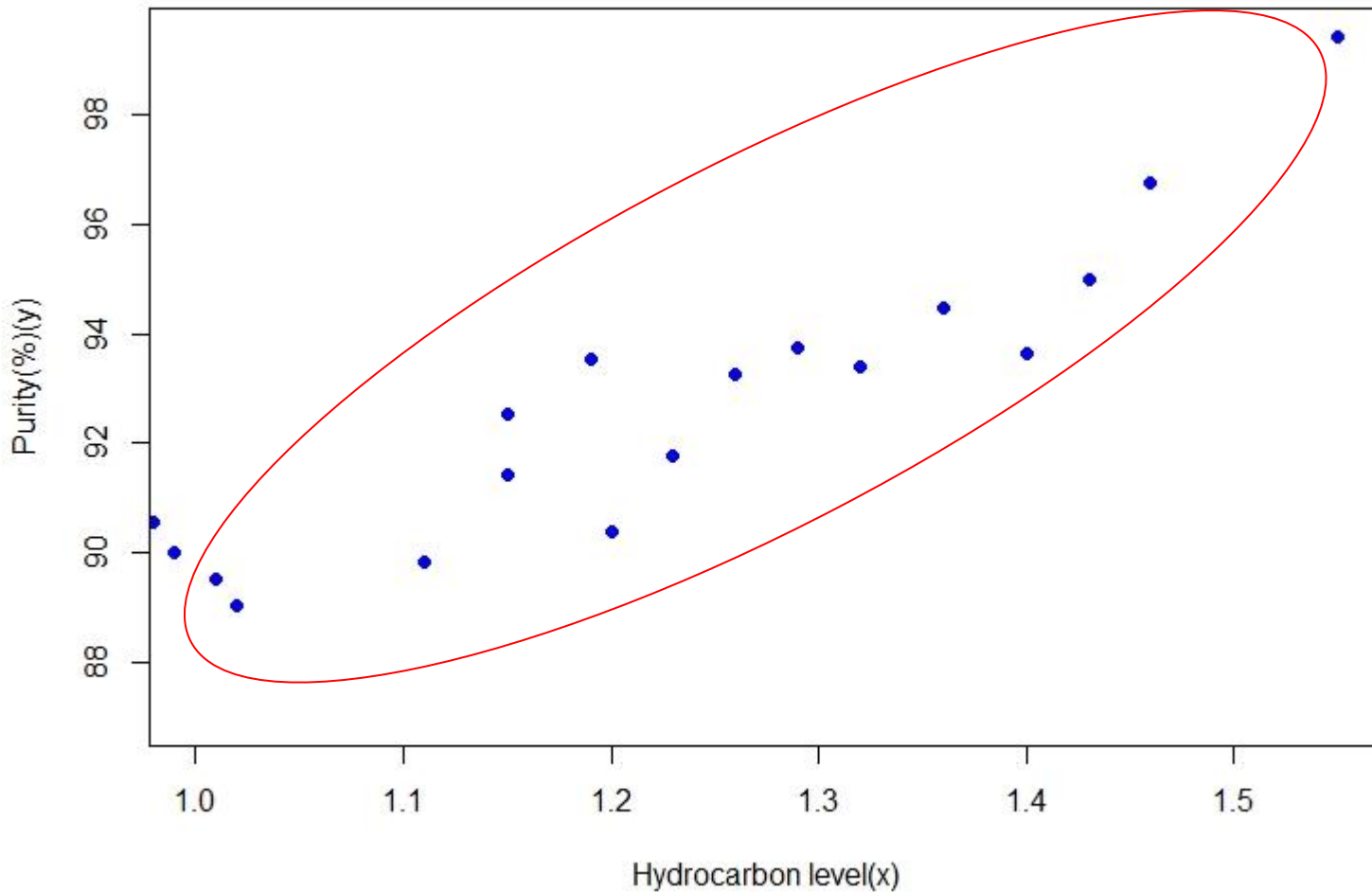


Figure 1. Scatter diagram of oxygen purity versus hydrocarbon level from Table 1.




Figure 1 presents a **scatter diagram**. Inspection of this scatter diagram indicates that, although no simple curve will pass exactly through all the points, there is a strong indication that the points lie scattered randomly around a straight line. Therefore, it is probably reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

where the slope (β_0) and intercept (β_1) of the line are called regression coefficients.

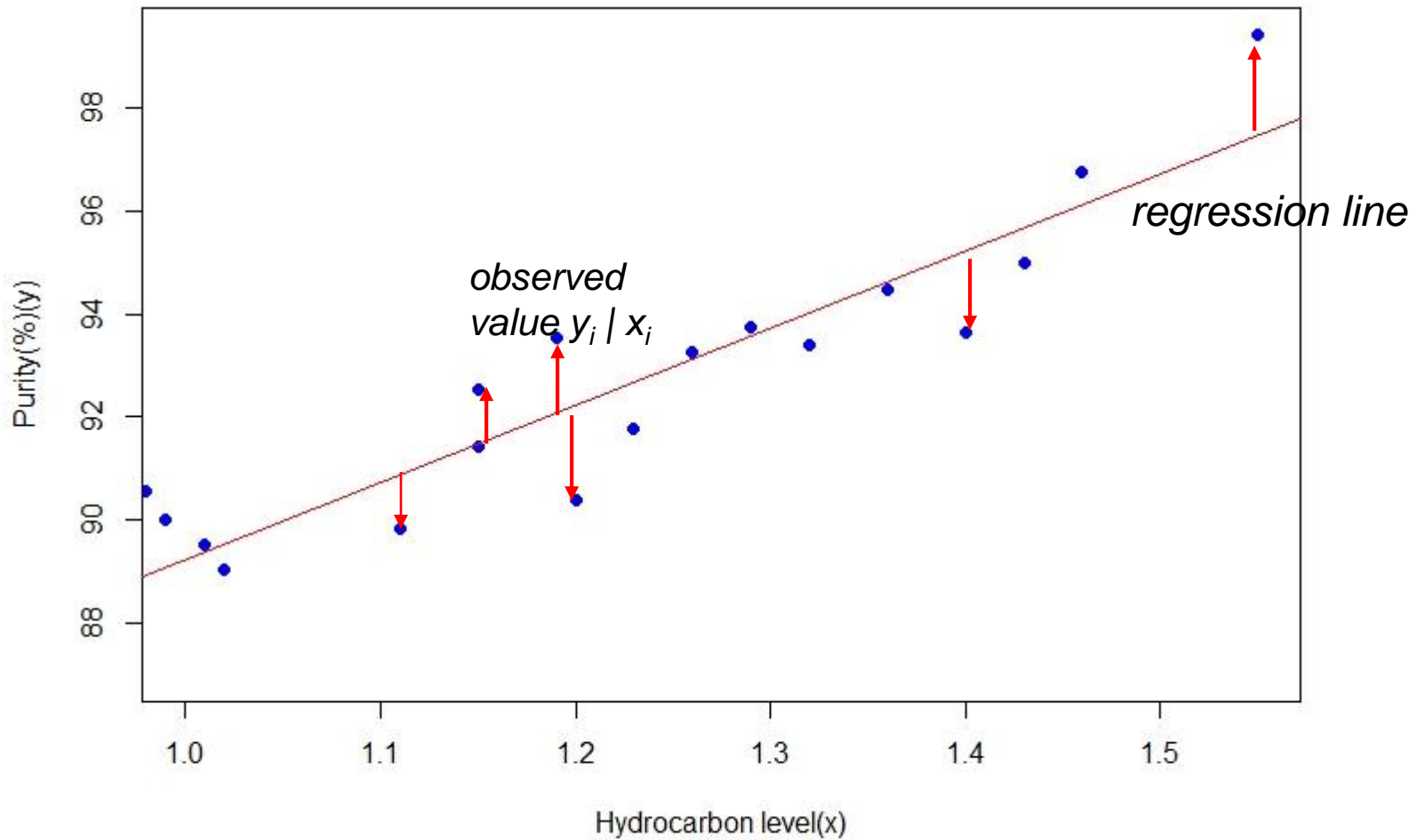


Figure 2. Deviations of the data from the estimated regression model.

We assume that each observation Y , can be described by the model

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

The diagram illustrates the components of the linear regression model equation $Y = \beta_0 + \beta_1 x + \varepsilon$. Three blue arrows point from the terms in the equation to labels in blue rounded rectangles below: β_0 points to *intercept*, β_1 points to *slope*, and ε points to *error*.

Where ε is a random error with mean zero and (unknown) variance σ_ε^2 . **The random errors corresponding to different observations are also assumed to be uncorrelated random variables.**

Suppose that we have n pairs of observations:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Figure 2 shows a typical scatter plot of observed data and a candidate for the estimated regression line. The estimates of β_0 and β_1 should result in a line that is (in some sense) a “best fit” to the data. The German scientist Karl Gauss (1777-1855) proposed estimating the parameters β_0 and β_1 in Equation (1) to minimize the sum of the squares of the vertical deviations in Figure 2.

We call this criterion for estimating the regression coefficients **the method of least squares.** Using Equation (1), we may express the n observations in the sample as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2)$$

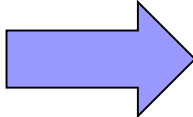
The sum of the squares of the deviations of the observations from the true regression line is:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (3)$$

The least squares estimators of β_0 and β_1 , say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\beta_0, \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\left. \frac{\partial L}{\partial \beta_1} \right|_{\beta_0, \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (4)$$

Simplifying these two equations yields

$$n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$
$$\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i$$


**Least squares
normal equations**

(5)

Definition: The least squares estimates of the intercept and slope in the simple linear regression model are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

The **fitted** or **estimated regression line** is therefore:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

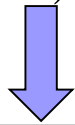
Note that each pair of observations satisfies the relationship

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \underbrace{e_i}_{y_i - \hat{y}_i}, \quad i = 1, 2, \dots, n$$

residual

Given data $(x_1, y_1)(x_2, y_2), \dots, (x_n, y_n)$

$$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$



Total Sum of Squares

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{S_{xy}}{SS_{xx}}$$

Example: We will fit a simple linear regression model to the oxygen purity data in Table 1. The following quantities may be computed:

$$n=20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2214.6566$$

$$SS_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088 \quad \text{and}$$

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2214.6566 - \frac{(23.92)(1843.21)}{20} = 10.17744$$

Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{S_{xy}}{SS_{xx}} = \frac{10.17744}{0.68088} = 14.94748 \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$

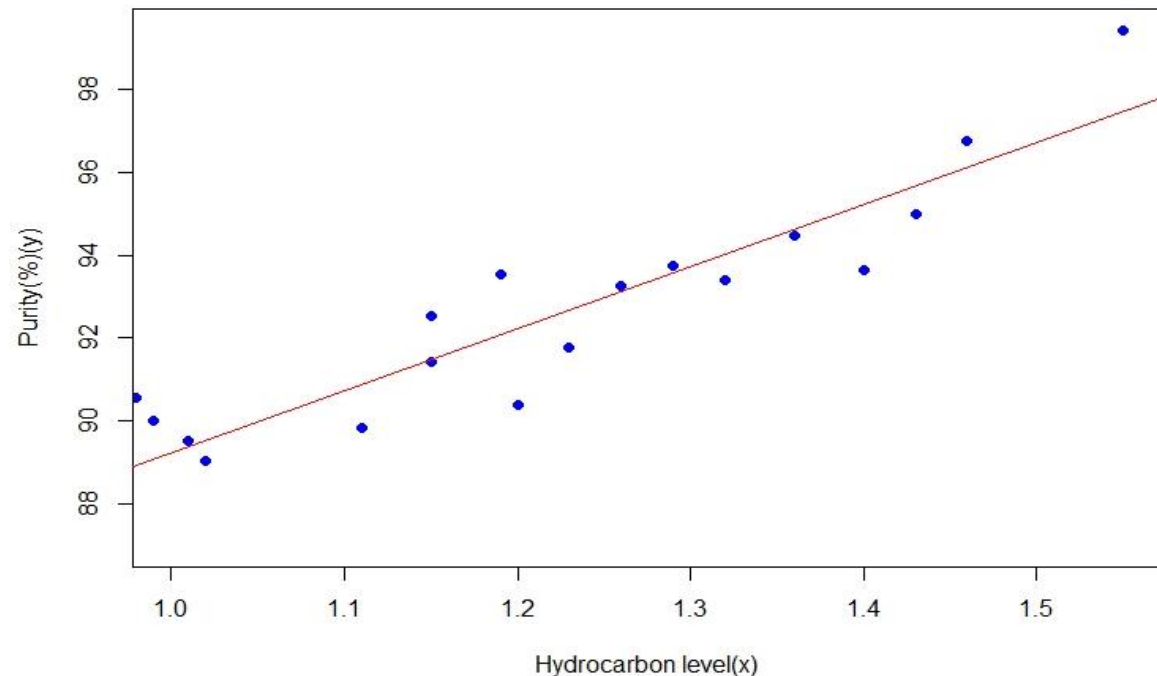



Figure 3. Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$


$$\hat{y} = 74.283 + 14.947x$$

Using the regression model of Example 1:

- we would predict oxygen purity of $\hat{y} = 89.23$ % when the hydrocarbon level is $x=1$ %. The purity 89.23 % may be interpreted as an estimate of the true population mean purity when $x=1.00$ %, or as an estimate of a new observation when $x=1.00$ %.
- $\hat{y} = 74.283$ % when the hydrocarbon level is $x=0$ %. The mean oxygen purity is 74.283 when there is no hydrocarbon level in process.

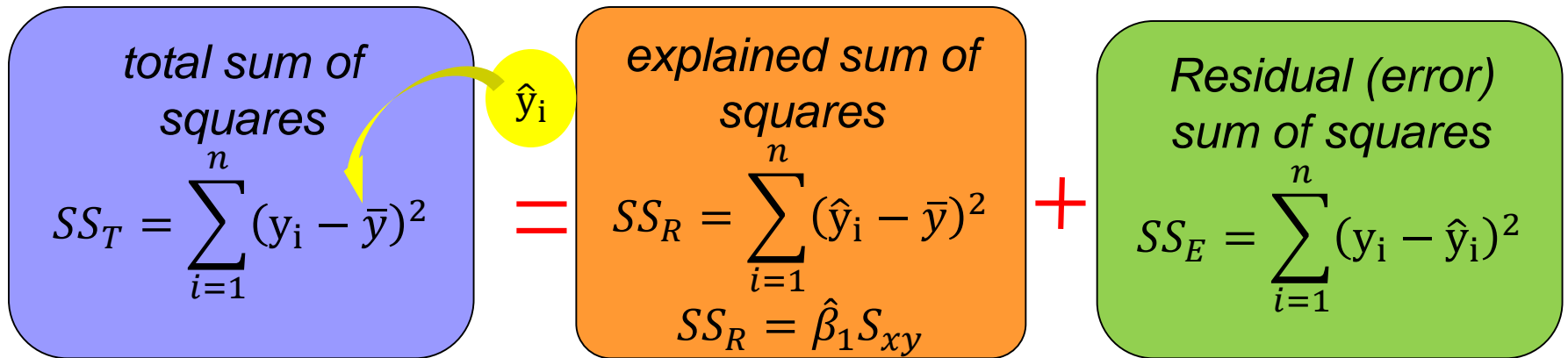
Estimating σ_ε^2

There is actually another unknown parameter in our regression model, σ_ε^2 (the variance of the error term ε). The residuals $e_i = y_i - \hat{y}_i$ are used to obtain an estimate of σ_ε^2 . The sum of squares of the residuals, often called the **error sum of squares**, is

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

We can show that the expected value of the error sum of squares is $E(SS_E) = (n - 2)\sigma_\varepsilon^2$. Therefore an unbiased estimator of σ_ε^2 is :

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{(n - 2)}$$



$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_T - \hat{\beta}_1 S_{xy}$$

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

the total sum of squares of the response variable y

The error sum of squares and the estimate of σ^2 for the oxygen purity data:

$$SS_T = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 170044.5321 - 20 \times (92.1605)^2 = 173.3769$$

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} = 173.3769 - 14.94748 \times 10.17744 = 21.249819$$

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n-2} = \frac{21.249819}{18} = 1.18$$

Properties of the Least Squares Estimators

The statistical properties of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ may be easily described. Recall that we have assumed that the error term ε in the model $Y = \beta_0 + \beta_1 x + \varepsilon$ is a random variable with mean zero and variance σ_ε^2 . Since the values of x are fixed, Y is a random variable with mean $\mu_{Y|x}$ and variance σ_ε^2 . Therefore, the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the observed y 's; thus, the least squares estimators of the regression coefficients may be viewed as random variables. We will investigate the bias and variance properties of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Consider first $\hat{\beta}_1$. Because $\hat{\beta}_1$ is a linear combination of the observations Y_i , we can use properties of expectation to show that expected value of $\hat{\beta}_1$ is:

$$E(\hat{\beta}_1) = \beta_1$$

$\hat{\beta}_1$ is an unbiased estimator of the true slope β_1 .

Since $V(\varepsilon_i) = \sigma_\varepsilon^2$ $V(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{SS_{xx}}$

For the intercept,

$$E(\hat{\beta}_0) = \beta_0 \quad V(\hat{\beta}_0) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]$$

$\hat{\beta}_0$ is an unbiased estimator of the intercept β_0 . The covariance of the random variables $\hat{\beta}_0$ and $\hat{\beta}_1$ is not zero. It can be shown that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma_\varepsilon^2 / SS_{xx}$.

Definition:

In simple linear regression the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}$$

respectively, where $\hat{\sigma}_\varepsilon^2$ is computed from the equation $\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n-2}$

Hypothesis Tests in Simple Linear Regression

An important part of assessing the adequacy of a linear regression model is testing statistical hypotheses about the model parameters and constructing certain confidence intervals. To test hypotheses about the slope and intercept of the regression model, we must make the additional assumption that the error component in the model, ε , is normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean zero and variance σ_ε^2 , abbreviated $N(0, \sigma_\varepsilon^2)$.

Use of t-Tests

Suppose we wish to test the hypothesis that the slope equals a constant, say, $\beta_{1,0}$. The appropriate hypotheses are :

$$H_0 : \beta_1 = \beta_1^*$$

$$H_1 : \beta_1 \neq \beta_1^*$$

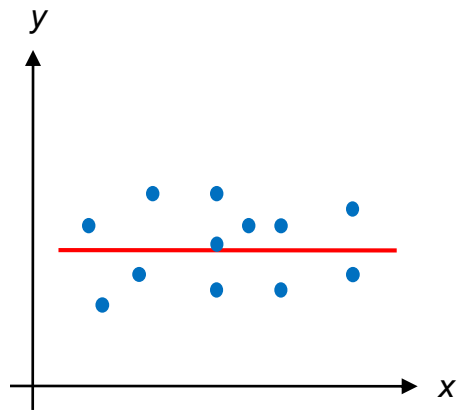
The other hypotheses test are:

$$H_0 : \beta_0 = \beta_0^*$$

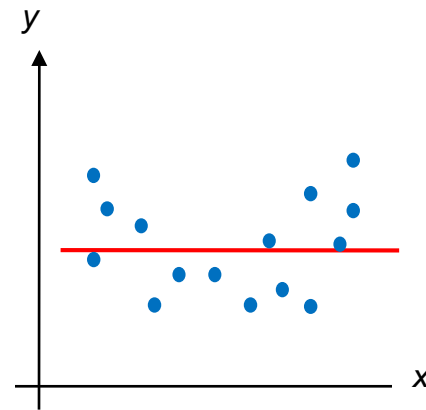
$$H_1 : \beta_0 \neq \beta_0^*$$

Since the errors ε_i are $N(0, \sigma_\varepsilon^2)$, it follows directly that the observations Y_i are $N(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$. Now $\hat{\beta}_1$ is a linear combination of independent normal random variables, and consequently, $\hat{\beta}_1$ is $N(\beta_1, \sigma_\varepsilon^2/SS_{xx})$. In addition, $(n-2)\hat{\sigma}_\varepsilon^2/\sigma_\varepsilon^2$ has a chi-square distribution with $n-2$ degrees of freedom $\left(\frac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi_{(n-2)}^2\right)$, and $\hat{\beta}_1$ is independent of $\hat{\sigma}_\varepsilon^2$. As a result of those properties, the statistic

Hypotheses	$H_0 : \beta_1 = \beta_1^*$ $H_1 : \beta_1 \neq \beta_1^*$	$H_0 : \beta_0 = \beta_0^*$ $H_1 : \beta_0 \neq \beta_0^*$
Test Statistic	$T = \frac{\hat{\beta}_1 - \beta_1^*}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}}}$	$T = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}}$
Decision	$ t > t_{\alpha/2, n-2}$ H_0 is rejected	$ t > t_{\alpha/2, n-2}$ H_0 is rejected

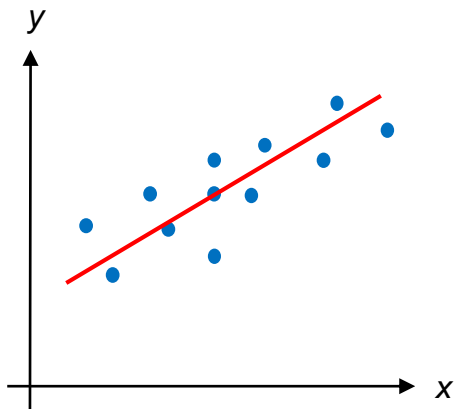


(a)

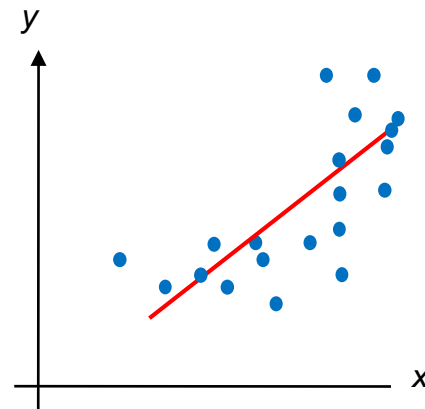


(b)

Figure 4. The hypothesis $H_0: \beta_1 = 0$ is not rejected.



(a)



(b)

Figure 5. The hypothesis $H_0: \beta_1 = 0$ is rejected.

Example: We will test for significance of regression using the model for the oxygen purity data. The hypotheses are

$$\begin{array}{l} H_0: \beta_1 = 0 \\ H_0: \beta_1 \neq 0 \end{array} \quad \alpha=0.05 \quad \hat{\beta}_1 = 14.97 \quad n = 20 \quad SS_{xx} = 0.68088 \quad \hat{\sigma}_\varepsilon^2 = 1.18$$

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}_\varepsilon^2 / SS_{xx}}} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18 / 0.68088}} = 11.35$$

test value $t = 11.35 \geq \underbrace{t_{0.025,18} = 2.101}_{\text{table value}}$ H_0 is rejected. Model parameter β_1 is significant at level 0.05.

$$\begin{array}{l} H_0: \beta_0 = 0 \\ H_0: \beta_0 \neq 0 \end{array} \quad t = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}} = \frac{\hat{\beta}_0}{\text{se}(\hat{\beta}_0)} = \frac{74.283}{\sqrt{1.18 \left[\frac{1}{20} + \frac{1.1960^2}{0.68088} \right]}} = 46.62$$

test value $t = 42.62 \geq \underbrace{t_{0.025,18} = 2.101}_{\text{table value}}$ H_0 is rejected. Model parameter β_0 is significant at level 0.05.

Analysis of Variance Approach to Test Significance of Regression

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SS_{total} (SS_T = SS_{yy}) = SS_{regression} (SS_R) + SS_{error} (SS_E)$$

H_0 : Model is not significant ($\beta_1 = 0$)

H_1 : Model is significant ($\beta_1 \neq 0$)

Table 2. Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	$MS_R = \frac{SS_R}{1}$	$f = \frac{MS_R}{MS_E}$
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	n-2	$MS_E = \frac{SS_E}{n-2}$	
Total	$SS_T = SS_{yy}$	n-1		

We reject H_0 hypothesis if $f \geq f_{\alpha, 1, n-2}$

Example: We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model. Recall that $SS_T = 173.38$, $\hat{\beta}_1 = 14.947$, $S_{xy} = 10.17744$, and $n=20$. The regression sum of squares is

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947) \times 10.17744 = 152.13$$

and the error sum of squares is

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

$$\begin{aligned} H_0: & \text{model is not significant } (\beta_1 = 0) \\ H_1: & \text{model is significant } (\beta_1 \neq 0) \end{aligned} \quad \alpha=0.05$$

Table 3. Analysis of Variance for Testing Significance of Regression for Oxygen Purity Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	152.13	1	152.13	128.92
Error	21.25 = SS_E	18	1.18 = $\hat{\sigma}_\varepsilon^2$	
Total	173.38	19		

Since $f = 128.92 \geq f_{0.05,1,18} = 4.414$, we reject H_0 at level $\alpha=0.05$.

**** Note** that the analysis of variance procedure for testing for significance of regression is equivalent to the *t*-test. That is, either procedure will lead to the same conclusions. This is easy to demonstrate by starting with the *t*-test statistic with $\beta_1^*=0$ say

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}_\varepsilon^2 / SS_{xx}}}$$

squaring both sides of the equation $T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}_\varepsilon^2 / SS_{xx}}}$ and using the fact that

$$MS_E = \hat{\sigma}_\varepsilon^2 \text{ results in } T^2 = \frac{\hat{\beta}_1^2 SS_{xx}}{MS_E} = \frac{\hat{\beta}_1 \left(\frac{\hat{\beta}_1}{SS_{xx}} \right) SS_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xy}}{MS_E} = \frac{MS_R}{MS_E}$$

T^2 is identical to F in the ANOVA table.

Confidence Intervals on the Slope and Intercept

In addition to point estimates of the slope and intercept, it is possible to obtain **confidence interval** estimates of these parameters. The width of these confidence intervals is a measure of the overall quality of the regression line. If the error term, ε_i , in the regression model are normally and independently distributed,

$$\left(\hat{\beta}_1 - \beta_1\right) / \sqrt{\hat{\sigma}_\varepsilon^2 / SS_{xx}} \quad \text{and} \quad \left(\hat{\beta}_0 - \beta_0\right) / \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}$$

are both distributed as t random variables with $n-2$ degrees of freedom. This leads to the following definition of $100(1-\alpha)\%$ confidence intervals on the slope and intercept.

Confidence Intervals on the Slope and Intercept

Definition: Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}}$$

Similarly, a $100(1-\alpha)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right]}$$

Example: We will find a 95% confidence interval on the slope of the regression line using the oxygen purity data. Recall that $\hat{\beta}_1 = 14.947$, $SS_{xx} = 0.68088$, and $\hat{\sigma}_\varepsilon^2 = 1.18$, then,

$$\hat{\beta}_1 - t_{0.025,18} \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18} \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{SS_{xx}}}$$

$$14.947 - 2.101 \sqrt{\frac{1.18}{0.68088}} \leq \beta_1 \leq 14.947 + 2.101 \sqrt{\frac{1.18}{0.68088}}$$

This simplifies to $12.197 \leq \beta_1 \leq 17.697$

Coefficient of Determination (R^2)

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad 0 \leq R^2 \leq 1$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model. Subsequently, in the case where X and Y are jointly distributed random variables, R^2 is the square of the correlation coefficient between X and Y. We often refer loosely to R^2 as the amount of variability in the data explained or accounted for by the regression model. For the oxygen purity regression model, we have $R^2 = \frac{SS_R}{SS_T} = \frac{152.13}{173.38} = 0.877$; that is, the model accounts for 87.7 % of the variability in the data.

For Example, suppose we wish to develop a regression model relating the shear strength of spot welds (noktasal kaynakların kopma/kesme mukavemeti) to the weld diameter (lehim, kaynak çapı). In this example, weld diameter cannot be controlled. We would randomly select n spot welds and observe a diameter and a shear strength (Y_i) for each. Therefore, (X_i, Y_i) are jointly distributed random variables. The sample correlation coefficient (R) between (X_i) and (Y_i) could be calculated as given in below:

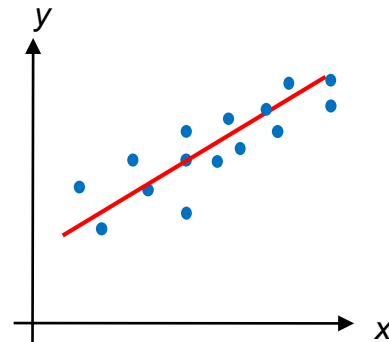
$$R = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2 \right]^{1/2}} = \frac{S_{xy}}{(SS_{xx}SS_T)^{1/2}}$$

$$-1 \leq R \leq 1$$

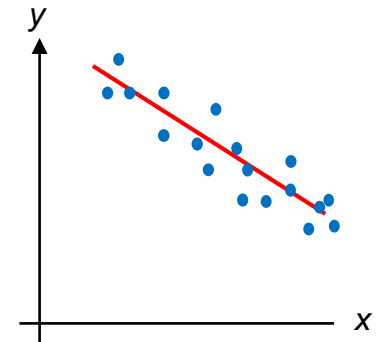
$R = 1$ shows strong positive linear relationship.

$R = 0$ shows no relationship.

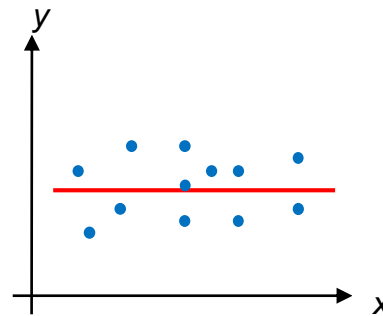
$R = -1$ shows strong negative linear relationship.



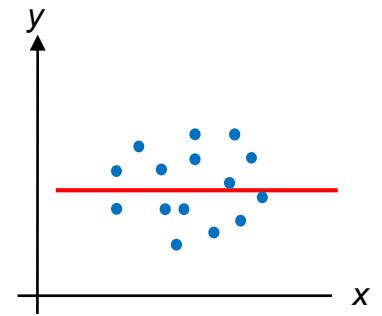
(a) $R \approx 1$



(b) $R \approx -1$



(c) $R \approx 0$



(d) $R \approx 0$

Figure 6: Correlation coefficient

Thus, $\hat{\beta}_1$ and R are closely related, although they provide somewhat different information. The sample correlation coefficient R measures the linear association between Y and X, while $\hat{\beta}_1$ measures the predicted change in the mean of Y for a unit change in X. In the case of a mathematical variable x, R has no meaning because the magnitude of R depends on the choice of spacing of x. We may also write:

$$R^2 = \hat{\beta}_1^2 \frac{SS_{xx}}{SS_{yy}} = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T}$$

which is just the coefficient of determination. That is, the coefficient of determination R^2 is just the square of the correlation coefficient between Y and X.

For the oxygen purity regression model, the sample correlation coefficient is

$$R = \frac{S_{xy}}{[SS_{xx}SS_T]^{1/2}} = \frac{10.17744}{[0.68088 \times 173.38]^{1/2}} = 0.9367 .$$