GOODNESS OF FIT

RESIDUALS

We used OLS method to develop an equation to describe the quantitative dependence between Y and X. Although the least squares method results in the line that fits the data with minimum distances, the regression equation is not a perfect predictor, unless all observed data points fall on the predicted regression line. We cannot expect all data points to fall exactly on the regression line. The regression line serves only as an approximate predictor of a Y value for a given value of X (or given values of X_1, X_2, …, X_k). Therefore, we need to develop a statistic that measures the variability of the actual values from the predicted Y values.

The differences between an observed Y value and the Y value predicted from the sample regression equation ($\hat{y}$) is called a residual.

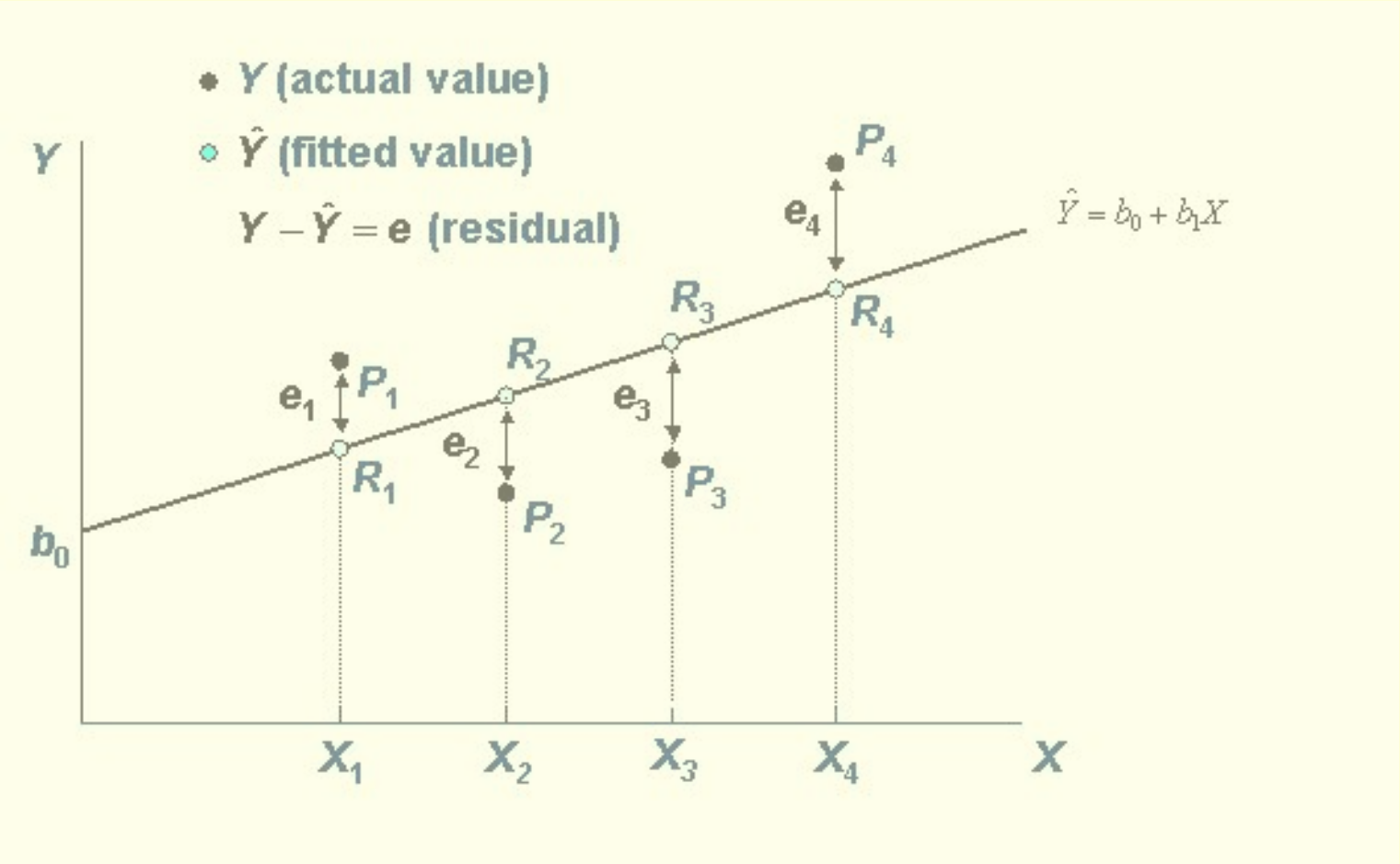
$$e_i = y_i - \hat{y}_i$$

residual for *i*-th observation

estimated value of the dependent variable using regression equation (simple or multiple) for *i*-th observation

actual value of Y for *i*-th observation

RESIDUALS



It should be emphasized that the residual is the vertical deviation of the observed Y value from the regression line.

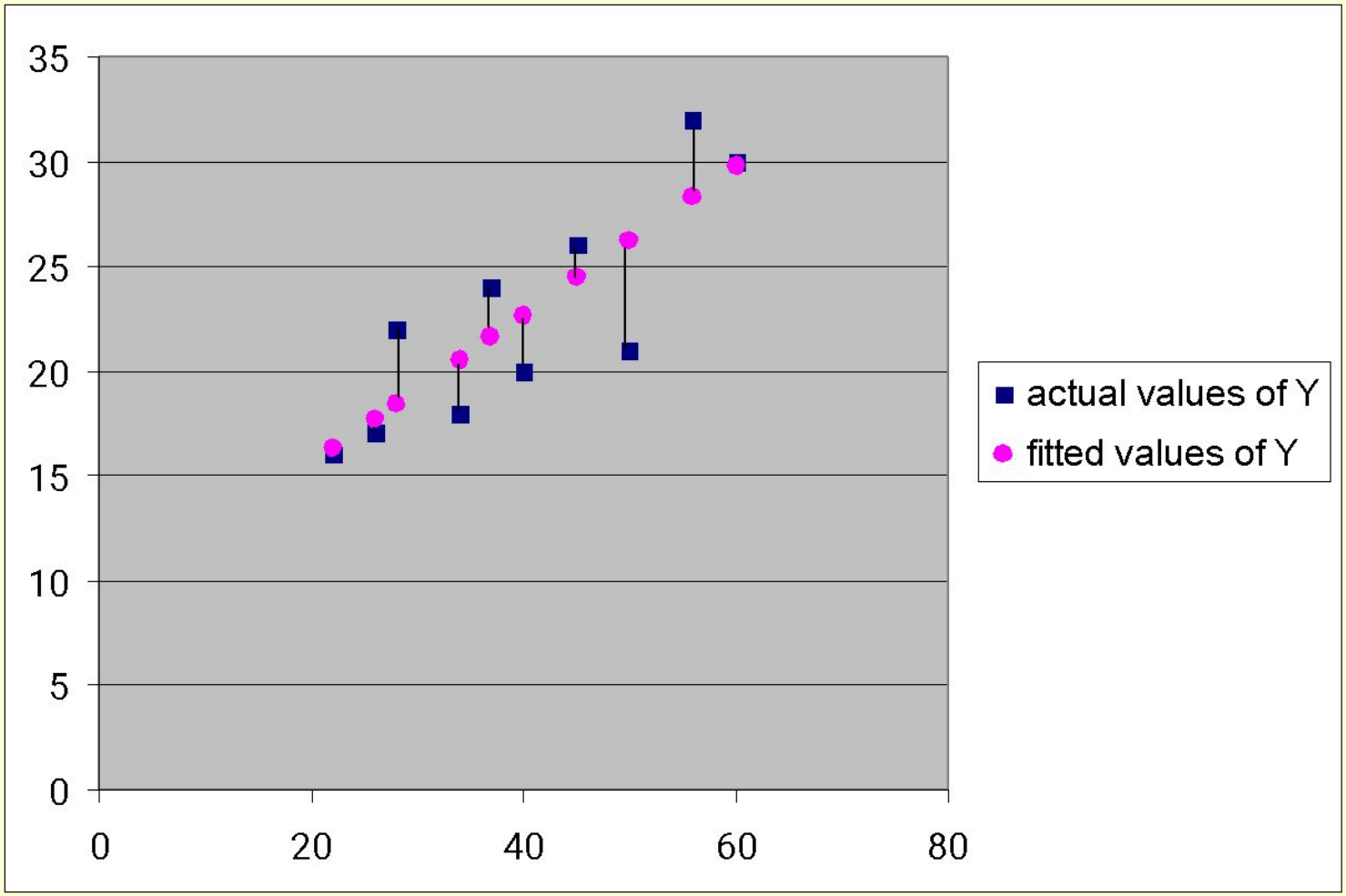
RESIDUALS

The $\hat{y}$ values are calculated by substituting the X value of each data pair into the regression equation.

Family	x_i	y_i	$\hat{y}_i = 8{,}51 + 0{,}35x_i$	$e_i = y_i - \hat{y}_i$
1	22	16	$\hat{y}_1$=8,51+0,35*22= 16,3	-0,30
2	26	17	$\hat{y}_2$=8,51+0,35*26= 17,7	-0,72
3	45	26	$\hat{y}_3$=8,51+0,35*45= 24,4	1,56
4	37	24	$\hat{y}_4$=8,51+0,35*37= 21,6	2,39
5	28	22	18,4	3,58
6	50	21	26,2	-5,21
7	56	32	28,3	3,67
8	34	18	20,5	-2,55
9	60	30	29,7	0,25
10	40	20	22,7	-2,67
		226	226	0,00

Estimators		
b_0	8,51	X - family income
b_1	0,35	Y - home size

RESIDUALS


35
30
25
20
15
10
5
0
0
20
40
60
80
actual values of Y
fitted values of Y

RESIDUALS

Y-weekly salary ($) **X1 –length of employment (months)** **X2-age (years)**

i	Y	X_1	X_2	$\hat{y}_i = 461{,}85 + 0{,}671 \cdot X_1 - 1{,}383 \cdot X_2$	$e_i = y_i - \hat{y}_i$
1	639	330	46	$\hat{y}_1$=461,85+0,671*330-1,383*46= 619,706	19,294
2	746	569	65	$\hat{y}_2$=461,85+0,671*569-1,383*65= 753,836	-7,836
3	670	375	57	$\hat{y}_3$=461,85+0,671*375-1,383*57= 634,692	35,308
4	518	113	47	472,674	45,326
5	602	215	41	549,436	52,564
6	612	343	59	610,447	1,553
7	548	252	45	568,736	-20,736
8	591	348	57	616,570	-25,570
9	552	352	55	622,021	-70,021
10	529	256	61	549,286	-20,286
11	456	87	28	481,508	-25,508
12	674	337	51	617,487	56,513
13	406	42	28	451,304	-45,304
14	529	129	37	497,247	31,753
15	528	216	46	543,190	-15,190
16	592	327	56	603,858	-11,858
	9192	**4291**	**779**	**9192**	0,000

b_0 461,85

b_1 0,671

b_2 -1,383

RESIDUALS

$$e_i = y_i - \hat{y}_i$$

The residual is the vertical deviation of the observed Y value from the regression surface.

STANDARD ERROR OF THE ESTIMATE

The measure of variability around the line of regression is called the standard error of the estimate (or estimation). It measures the typical difference between the actual values and the Y values predicted by the regression equation. This can be seen by the formula for the standard error of the estimate:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k - 1}}$$

standard error of the estimate

sample Y values

values of Y calculated from the regression equation

number of predictors

sample size

It is measured in units of the dependent variable Y.

STANDARD ERROR OF THE ESTIMATE IS A MEASURE OF THE VARIABILITY, OR SCATTER, OF THE OBSERVED SAMPLE Y VALUES AROUND THE REGRESSION LINE.

STANDARD ERROR OF THE ESTIMATE

Let's calculate standard error of estimation for our simple regression equation (X – family income, Y – home size. If you are lost, see slide no. 4)

b_0= 8,51

b_1= 0,35

Family	x_i	y_i	y^	$e_i=y_i - y^_i$	e_i^2
1	22	16	16,3	-0,30	0,09
2	26	17	17,716	-0,72	0,51
3	45	26	24,44	1,56	2,43
4	37	24	21,609	2,39	5,72
5	28	22	18,424	3,58	12,79
6	50	21	26,21	-5,21	27,14
7	56	32	28,334	3,67	13,44
8	34	18	20,547	-2,55	6,49
9	60	30	29,749	0,25	0,06
10	40	20	22,671	-2,67	7,13
		226	226	0,00	**75,81**

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{75,81}{10-1-1}} = \sqrt{9,48} = 3,08$$

STANDARD ERROR OF THE ESTIMATE

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{75{,}81}{10-1-1}} = \sqrt{9{,}48} = 3{,}08$$

What does it mean?

To answer this question, you must refer to the units in which the Y variable is measured.

Home size is measured in hundreds of square feet.

THE ACTUAL VALUES OF HOME SIZE DIFFER FROM THE ESTIMATED VALUES (USING REGRESSION EQUATION) OF HOME SIZE FOR 308 SQUARE FEET, ON AVERAGE.

STANDARD ERROR OF THE ESTIMATE

Let's calculate standard error of estimation for our multiple regression equation

Y-weekly salary ($) X1 –length of employment (months) X2-age (years)

(if you are lost, see slide no. 6)

b_0= 461,85

b_1= 0,671

b_2= -1,383

i	Y	X_1	X_2	y^	$e_i=y_i - y^_i$	e_i^2
1	639	330	46	619,706	19,294	372,254
2	746	569	65	753,836	-7,836	61,405
3	670	375	57	634,692	35,308	1246,651
4	518	113	47	472,674	45,326	2054,471
5	602	215	41	549,436	52,564	2762,970
6	612	343	59	610,447	1,553	2,412
7	548	252	45	568,736	-20,736	430,001
8	591	348	57	616,570	-25,570	653,817
9	552	352	55	622,021	-70,021	4903,007
10	529	256	61	549,286	-20,286	411,535
11	456	87	28	481,508	-25,508	650,653
12	674	337	51	617,487	56,513	3193,685
13	406	42	28	451,304	-45,304	2052,471
14	529	129	37	497,247	31,753	1008,244
15	528	216	46	543,190	-15,190	230,738
16	592	327	56	603,858	-11,858	140,617
	9192	**4291**	**779**	**9192**	0,000	**20174,9311**

STANDARD ERROR OF THE ESTIMATE

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{20174{,}9311}{16-2-1}} = \sqrt{\frac{1551{,}9178}{13}} = 39{,}394$$

What does it mean?

To answer this question, you must refer to the units in which the Y variable is measured.

Variable Y is weekly salary. Its unit is $.

THE ACTUAL VALUES OF WEEKLY SALARY DIFFER FROM THE ESTIMATED VALUES (USING REGRESSION EQUATION) FOR 39,39 $, ON AVERAGE.

THE MEAN DIFFERENCES BETWEEN THE ACTUAL AND PREDICTED VALUES OF WEEKLY SALARY ARE EQUAL 39,39 $, ON AVARAGE.

COEFFICIENT OF RESIDUAL'S VARIABILITY

Coefficient of residual variability measures a percent of standard error of the estimate from the mean Y value. Its unit is %. We calculate it using formula:

$$V_e = \frac{S_e}{\bar{y}} \cdot 100$$

Good model is a regression model with V_e lower than 15%.

COEFFICIENT OF RESIDUAL'S VARIABILITY

For our examples:

X – family income

Y – home size

Se = 3,08 [hundreds of square feet]

$$\bar{y} = \frac{226}{10} = 22{,}6$$

$$V_e = \frac{S_e}{\bar{y}} \cdot 100 = \frac{3{,}08}{22{,}6} \cdot 100 = 13{,}63\%$$

X1– length of employment

X2 – age

Y – weekly salary

Se = 39,394 [$]

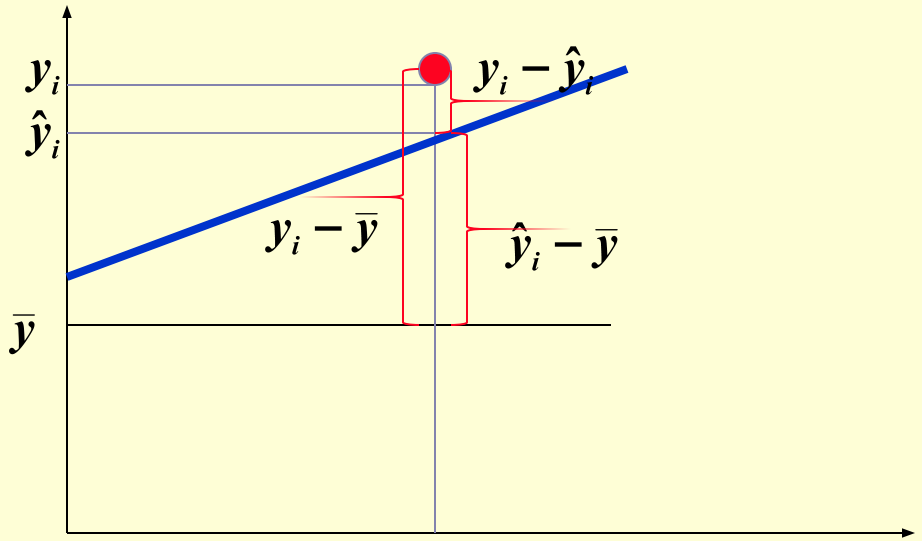
$$\bar{y} = \frac{9192}{16} = 574{,}5$$

$$V_e = \frac{S_e}{\bar{y}} \cdot 100 = \frac{39{,}394}{574{,}5} \cdot 100 = 6{,}86\%$$

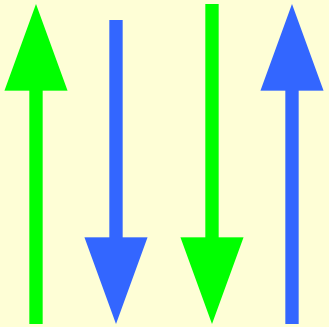
HOW GOOD IS OUR MODEL?

In order to examine how well the independent variable (or variables) predicts the dependent variable in our model, we need to develop several measures of variation. The first measure, the TOTAL SUM OF SQUARES (SST), is a measure of variation (or scatter) of the Y values around the mean. The total sum of squares can be subdivided into explained variation (or REGRESSION SUM OF SQUARES, SSR), that is attributable to the relationship between the independent variable (or variables) and the dependent variable, and unexplained variation (or ERROR SUM OF SQUARES, SSE), that which is attributable to factors other than the relationship between the independent variable (or variables) and the dependent variable.

HOW GOOD IS OUR MODEL?



SST= SSR + SSE



$\sum (y_i - \bar{y})^2$ **=SST (TOTAL SUM OF SQUARES)**

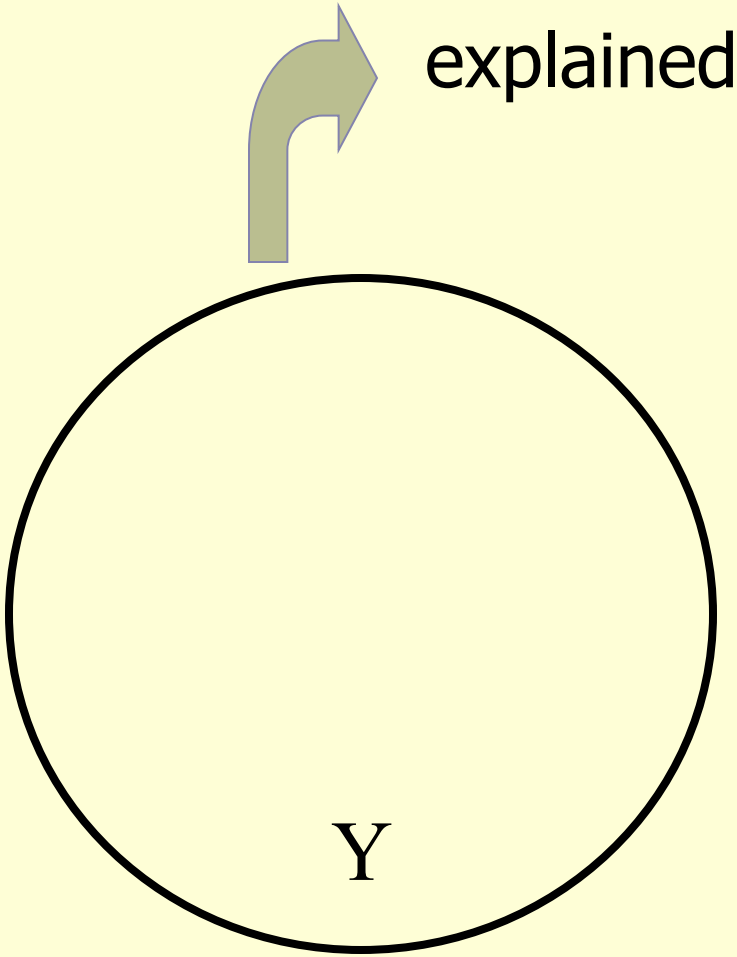
$\sum (\hat{y}_i - \bar{y})^2$ **=SSR (EXPLAINED SUM OF SQUARES)**

$\sum (y_i - \hat{y}_i)^2$ **=SSE (UNEXPLAINED SUM OF SQUARES)**

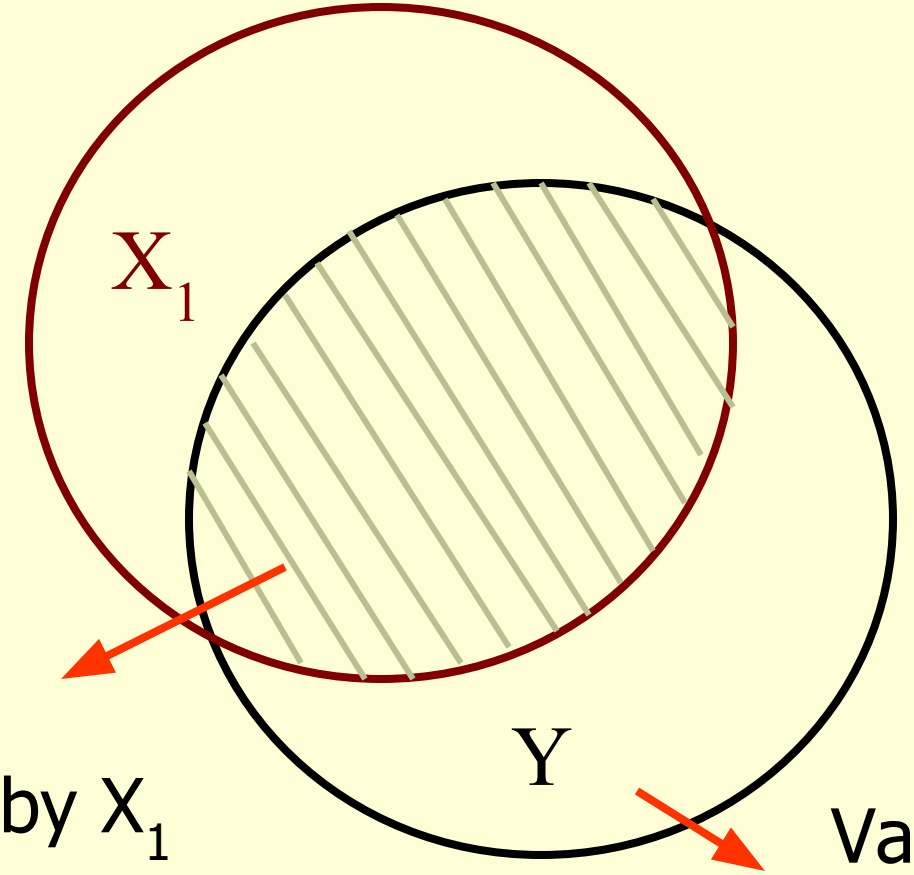
$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

HOW GOOD IS OUR MODEL?

Variance to be explained by predictors



HOW GOOD IS OUR MODEL?



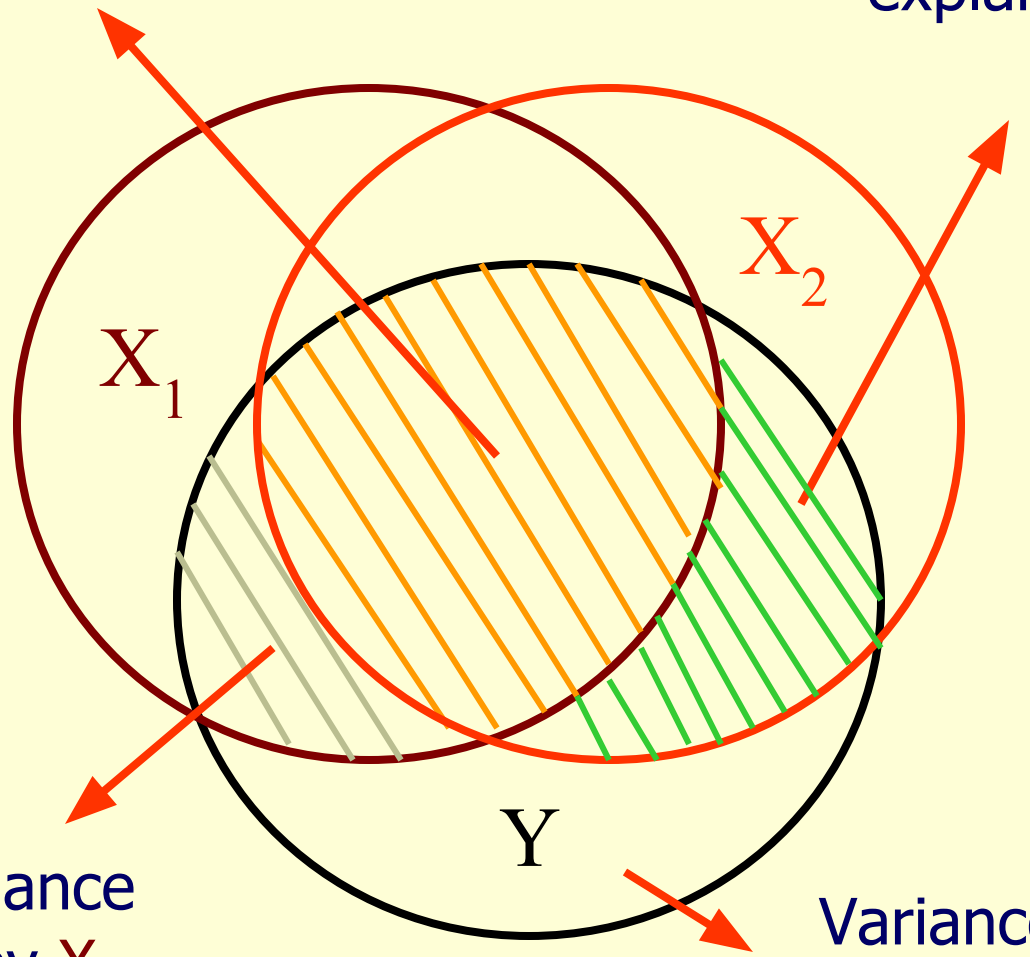
Variance explained by X_1

Variance NOT explained by X_1

HOW GOOD IS OUR MODEL?

Common variance explained by X_1 and X_2

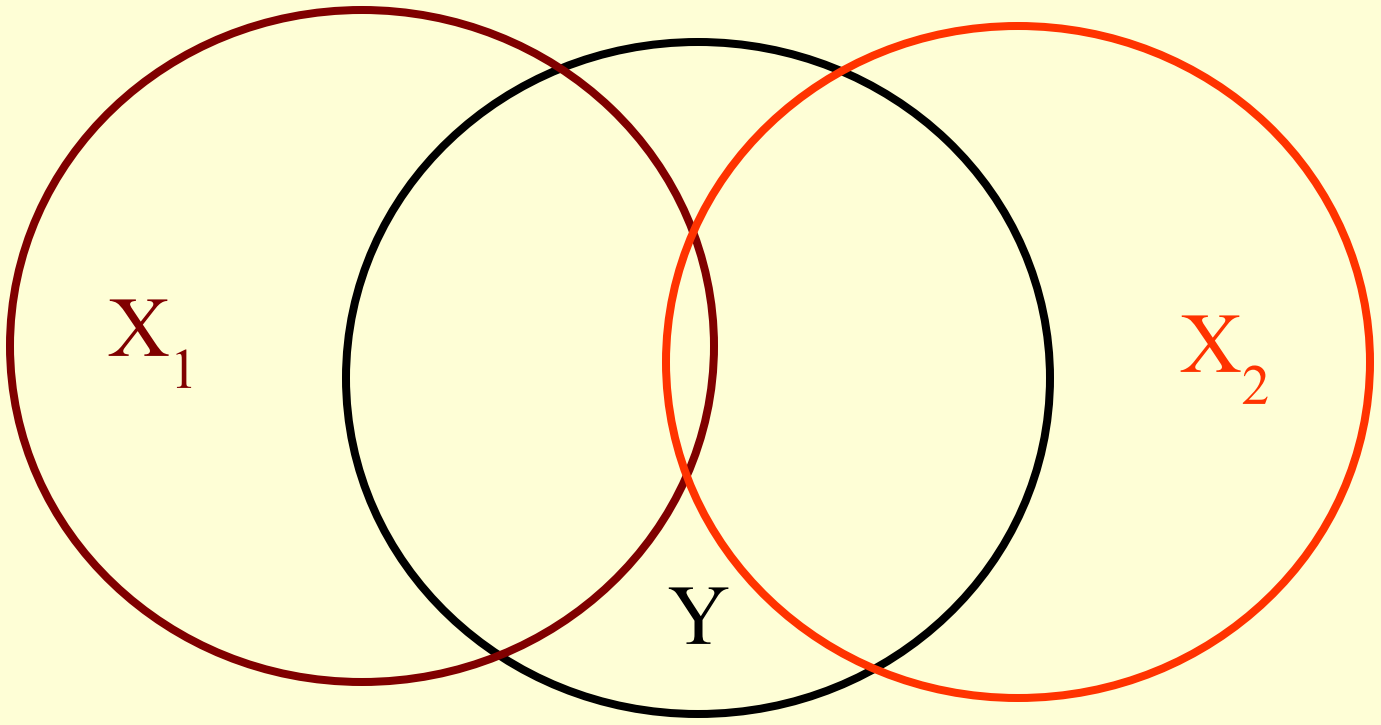
Unique variance explained by X_2



Unique variance explained by X_1

Variance NOT explained by X_1 and X_2

A “good” model



DETERMINATION COEFFICIENT

The coefficient of determination, R^2, of the fitted regression is defined as the proportion of the total sample variability explained by the regression and is

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

and it follows that

$$0 \leq R^2 \leq 1$$

R^2 gives the proportion of the total variation in the dependent variable explained by the independent variable (or variables).

If R^2 = 1, then ???

If R^2 = 0, then ???

INDETERMINATION COEFFICIENT

The coefficient of indetermination, φ^2, of the fitted regression is defined as the proportion of the total sample variability unexplained by the regression and is

$$\varphi^2 = \frac{SSE}{SST}$$

and it follows that

$$0 \leq \varphi^2 \leq 1$$

φ^2 **gives the proportion of the total variation in the dependent variable unexplained by the independent variable (or variables).**

If it's equal to 1, then ???

If it's equal to 0, then ???

$$\varphi^2 + R^2 = 1$$

ADJUSTED COEFFICIENT OF DETERMINATION

The adjusted coefficient of determination, R^2, is defined as

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

or

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

We use this measure to correct for the fact that non-relevant independent variables will result in some small reduction in the error sum of squares. Thus the adjusted R^2 provides a better comparison between multiple regression models with different numbers of independent variables. Since R^2 always increases with the addition of a new variable, the adjusted R^2 compensates for added explanatory variables.

COEFFICIENT OF MULTIPLE CORRELATION

The coefficient of multiple correlation, is the correlation between the predicted value and the observed value of the dependent variable:

$$R = Corr(\hat{Y}, y) = \sqrt{R^2}$$

and is equal to the square root of the coefficient of determination.

We use R as another measure of the strength of the linear relationship between the dependent variable and the independent variable (or variables). Thus it is comparable to the correlation between Y and X in simple regression.

$$0 \leq R \leq 1$$

DETERMINATION COEFFICIENT – EXAMPLE – ONE REGRESSOR

Let's calculate coefficient of determination (and indetermination) for our multiple regression equation (slide no. 4 and 9)

Y-home size X –family income

b_0= 8,51

b_1= 0,35

Family	x_i	y_i	y^	$e_i=y_i - y^_i$	e_i^2	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	22	16	16,3	-0,30	0,09	-6,6	43,56
2	26	17	17,716	-0,72	0,51	-5,6	31,36
3	45	26	24,44	1,56	2,43	3,4	11,56
4	37	24	21,609	2,39	5,72	1,4	1,96
5	28	22	18,424	3,58	12,79	-0,6	0,36
6	50	21	26,21	-5,21	27,14	-1,6	2,56
7	56	32	28,334	3,67	13,44	9,4	88,36
8	34	18	20,547	-2,55	6,49	-4,6	21,16
9	60	30	29,749	0,25	0,06	7,4	54,76
10	40	20	22,671	-2,67	7,13	-2,6	6,76
		226	226	0,00	**75,81**		**262,4**

DETERMINATION COEFFICIENT – EXAMPLE – ONE REGRESSOR

The coefficient of determination should be calculated as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{75,81}{262,4} = 1 - 0,29 = 0,71$$

It's easy to provide the coefficient of indetermination:

$$\varphi^2 = \frac{SSE}{SST}$$

IT CAN BE SAID THAT 29% OF THE VARIABILITY IN HOME SIZES (Y) REMAINS UNEXPLAINED BY THE FAMILY INCOME. THEREFORE, 71% OF THE VARIABILITY IN HOME SIZES (Y) IS EXPLAINED BY THE PREDICTOR.

WE HAVE ACCOUNTED FOR 71% OF THE TOTAL VARIATION IN THE HOME SIZES BY USING INCOME AS A PREDICTOR OF HOME SIZE.

DETERMINATION COEFFICIENT – EXAMPLE – TWO REGRESSORS

Let's calculate coefficient of determination (and indetermination) for our multiple regression equation (slide no. 6 and 11)

Y-weekly salary ($) X1 –length of employment (months) X2-age (years)

$b_0 = 461{,}85$

$b_1 = 0{,}671$

$b_2 = -1{,}383$

$$\bar{y} = \frac{9192}{16}$$

$$\bar{y} = 574{,}5$$

i	Y	X_1	X_2	y^	$e_i = y_i - y^\wedge_i$	e_i^2	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	639	330	46	619,706	19,294	372,254	**64,5**	**4160,25**
2	746	569	65	753,836	-7,836	61,405	**171,5**	**29412,25**
3	670	375	57	634,692	35,308	1246,651	**95,5**	**9120,25**
4	518	113	47	472,674	45,326	2054,471	**-56,5**	**3192,25**
5	602	215	41	549,436	52,564	2762,970	**27,5**	**756,25**
6	612	343	59	610,447	1,553	2,412	**37,5**	**1406,25**
7	548	252	45	568,736	-20,736	430,001	**-26,5**	**702,25**
8	591	348	57	616,570	-25,570	653,817	**16,5**	**272,25**
9	552	352	55	622,021	-70,021	4903,007	**-22,5**	**506,25**
10	529	256	61	549,286	-20,286	411,535	**-45,5**	**2070,25**
11	456	87	28	481,508	-25,508	650,653	**-118,5**	**14042,25**
12	674	337	51	617,487	56,513	3193,685	**99,5**	**9900,25**
13	406	42	28	451,304	-45,304	2052,471	**-168,5**	**28392,25**
14	529	129	37	497,247	31,753	1008,244	**-45,5**	**2070,25**
15	528	216	46	543,190	-15,190	230,738	**-46,5**	**2162,25**
16	592	327	56	603,858	-11,858	140,617	**17,5**	**306,25**
	9192	**4291**	**779**	**9192**	0,000	**20174,9311**		**108472**

DETERMINATION COEFFICIENT – EXAMPLE – TWO REGRESSORS

The coefficient of determination should be calculated as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{20174{,}9311}{108472{,}00} = 1 - 0{,}186 = 0{,}814$$

It's easy to provide the coefficient of indetermination:

$$\varphi^2 = \frac{SSE}{SST}$$

IT CAN BE SAID THAT 18,6% OF THE VARIABILITY IN WEEKLY SALARY (Y) REMAINS UNEXPLAINED BY LENGTH OF EMPLOYMENT (X1) AND THE AGE (X2) OF EMPLOYEES. THEREFORE, 81,4% OF THE VARIABILITY IN WEEKLY SALARY (Y) IS EXPLAINED BY THESE TWO PREDICTORS.

ADJUSTED COEFFICIENT OF DETERMINATION - EXAMPLE

We can compare these two models using adjusted coefficient of determination.

For regression model with one regressor (see slide 26) :

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{75{,}81/(10-1-1)}{262{,}4/9} =$$

$$= 1 - \frac{9{,}48}{29{,}1} = 1 - 0{,}326 = 0{,}674$$

For regression model with two predictors (see slide 28):

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{16-1}{16-2-1}(1-0{,}814) =$$

$$= 1 - 0{,}215 = 0{,}785$$

This is better result of goodness of fit.

COEFFICIENT OF MULTIPLE CORRELATION

The coefficient of multiple correlation, is the square root of the multiple coefficient of determination:

$$R = \sqrt{R^2}$$

For regression model with 1 independent variable (Y – home size, X – family income, R^2=0,71; see slide no.26)

$$R = \sqrt{R^2} = \sqrt{0,71} = 0,843$$

There is strong positive correlation between the home size and the family income.

For regression model with 2 independent variables (Y – salary, X1 – length of employment,, X2 – age, R^2=0,814; see slide no.28)

$$R = \sqrt{R^2} = \sqrt{0,814} = 0,902$$

There is very strong correlation between the salary and the length of employment and the age.