Linear Regression

DR.ALAA MOHAMMED

3RD LEVEL, 2ND SEMESTER, BIOMEDICAL INSTRUMENTATION AND BIOMECHANICS BRANCHES, BIOMEDICAL ENG. DEPARTMENT

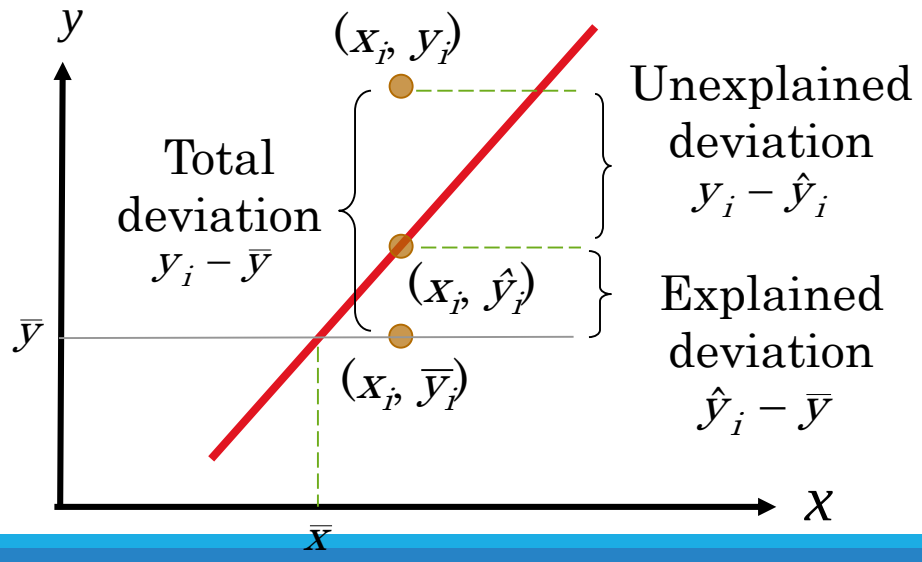
Measures of Regression and Prediction Intervals

PART3

Variation About a Regression Line

To find the total variation, you must first calculate the **total deviation**, the **explained deviation**, and the **unexplained deviation**.

- Total deviation $= y_i - \bar{y}$
- Explained deviation $= \hat{y}_i - \bar{y}$
- Unexplained deviation $= y_i - \hat{y}_i$



Variation About a Regression Line (cont.)

The **total variation** (total sum of squares) about a regression line is the sum of the squares of the differences between the *y*-value of each ordered pair and the mean of *y*.

$$\textbf{Total variation} = \sum(y_i - \bar{y})^2$$

The **explained variation** (explained sum of squares) is the sum of the squares of the differences between each predicted *y*-value and the mean of *y*.

$$\textbf{Explained variation} = \sum(\hat{y}_i - \bar{y})^2$$

The **unexplained variation** (Residual sum of squares) is the sum of the squares of the differences between the *y*-value of each ordered pair and each corresponding predicted *y*-value.

Total variation = Explained variation + Unexplained variation

$$\textbf{Unexplained variation} = \sum(y_i - \hat{y}_i)^2$$

Coefficient of Determination R

The **coefficient of determination r^2** is the ratio of the explained variation to the total variation. That is,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Example 2 (cont.):

The correlation coefficient for the data that represents the number of hours students watched television and the test scores of each student is $r \approx -0.831$.

Find the coefficient of determination.

R= $r^2 \approx (-0.831)^2$

≈ 0.691

About 69.1% of the variation in the test scores can be explained by the variation in the hours of TV watched. About 30.9% of the variation is unexplained.

The Standard Error of Estimate

When a $\hat{y}$-value is predicted from an *x*-value, the prediction is a point estimate.

An interval can also be constructed

The **standard error of estimate s_e** is the standard deviation of the observed y_i -values about the predicted $\hat{y}$-value for a given x_i -value. It is given by

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

where ***n*** is the number of ordered pairs in the data set

The closer the observed *y*-values are the predicted *y*-values, the smaller the standard error of estimate will be.

The Standard Error of Estimate (cont.)

As the **standard error of the estimate** (the variability of the data about the regression line) rises, the confidence widens. In other words, the more variable the data, the less confident you will be when you're using the regression model to estimate the coefficient

The Standard Error of Estimate (Procedure)

Finding the Standard Error of Estimate

1. Make a table that includes the column reading shown. $x_i, y_i, \hat{y}_i, (y_i - \hat{y}_i), (y_i - \hat{y}_i)^2$
2. Use the regression equation to calculate the predicted *y*-values. $\hat{y} = mx_i + b$
3. Calculate the sum of the squares of the differences between each observed *y*-value and the corresponding predicted *y*-value. $\sum(y_i - \hat{y}_i)^2$
4. Find the standard error of estimate.

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}$$

Example 4:

The regression equation for the following data is $\hat{y} = 1.2x - 3.8$.
Find the standard error of estimate.

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	– 3	– 2.6	0.16
2	– 1	– 1.4	0.16
3	0	– 0.2	0.04
4	1	1	0
5	2	2.2	0.04
			$\sum = 0.4$

Unexplained variation

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{0.4}{5-2}} \approx 0.365$$

The standard deviation of the predicted *y*_value for a given *x* value is about 0.365.

Example 5:

The regression equation for the data that represents the number of hours **12** different students watched television during the weekend and the scores of each student who took a test the following Monday is

$$\hat{y} = -4.07x + 93.97$$

Find the standard error of estimate.

Hours, x_i	0	1	2	3	3	5
Test score, y_i	96	85	82	74	95	68
$\hat{y}_i$	93.97	89.9	85.83	81.76	81.76	73.62
$(y_i - \hat{y}_i)^2$	4.12	24.01	14.67	60.22	175.3	31.58

Continued.

Example 5 (cont.)

Hours, x_i	5	5	6	7	7	10
Test score, y_i	76	84	58	65	75	50
$\hat{y}_i$	73.62	73.62	69.55	65.48	65.48	53.27
$(y_i - \hat{y}_i)^2$	5.66	107.74	133.4	0.23	90.63	10.69

$\sum(y_i - \hat{y}_i)^2 = 658.25$

Unexplained variation

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

The standard deviation of the student test scores for a specific number of hours of TV watched is about 8.11

Prediction Intervals

Two variables have a **bivariate normal distribution** if for any fixed value of $\boldsymbol{x}$, the

corresponding values of $\boldsymbol{y}$ are normally distributed and for any fixed values of $\boldsymbol{y}$, the corresponding x-values are normally distributed.

A prediction interval can be constructed for the true value of y.

Given a linear regression equation $\hat{y} = mx + b$ and x_0, a specific value of x, a ***c*-prediction interval** for y is

$$\hat{y} - \boldsymbol{E} < \boldsymbol{y} < \hat{y} + \boldsymbol{E}$$

where

$$\boldsymbol{E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}}$$

The point estimate is $\hat{y}$ and the margin of error is $\boldsymbol{E}$. The probability that the prediction interval contains $\boldsymbol{y}$ is $\boldsymbol{c}$.

Prediction Intervals (cont.)

Construct a Prediction Interval for y for a Specific Value of x

1. Identify the number of ordered pairs in the data set n and the degrees of freedom.

$$\text{d.f.} = n - 2$$

2. Use the regression equation and the given x-value to find the point estimate ŷ.

$$\hat{y} = mx_i + b$$

3. Find the critical value t_c that corresponds to the given level of confidence c.

Use Table

Prediction Intervals (cont.)

4. Find the standard error of estimate s_e.

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

5. Find the margin of error ***E***.

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$$

6. Find the left and right endpoints and form the prediction interval.

Left endpoint: $\hat{y} - E$
Right endpoint: $\hat{y} + E$
Interval: $\hat{y} - E < y < \hat{y} + E$

Example 3:

The following data represents the number of hours **12** different students watched television during the weekend and the scores of each student who took a test the following Monday.

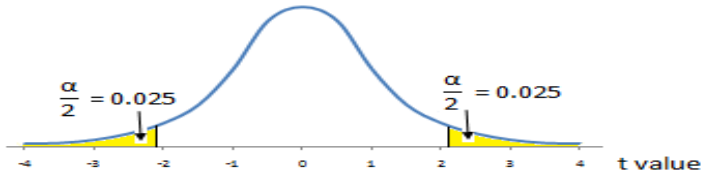
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

$$\hat{y} = -4.07x + 93.97 \qquad s_e \approx 8.11$$

Construct a 95% prediction interval for the test scores when 4 hours of TV are watched.

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	

Example 3 (cont.)

Construct a 95% prediction interval for the test scores when the number of hours of TV watched is 4.

There are $n - 2 = 12 - 2 = 10$ degrees of freedom

The point estimate is

$$\hat{y} = -4.07x + 93.97 = -4.07(4) + 93.97 = 77.69.$$

The critical value $t_c = 2.228$, and $s_e = 8.11$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}} \quad ,$$

$$\boldsymbol{E = (2.228)(8.11)\sqrt{1 + \frac{1}{12} + \frac{12(4-(4.5))}{12(332)-(54)^2}} \approx 18.75}$$

$$\hat{y} - E < y < \hat{y} + E \qquad 77.69 - 18.75 = 58.94 \qquad , 77.69 + 18.75 = 96.44$$

You can be 95% confident that when a student watches 4 hours of TV over the weekend, the student's test grade will be between 58.94 and 96.44.