Section 4.7 – Confidence and Prediction Intervals

Method for constructing a prediction interval, which is an interval estimate of a predicted value of ***y***

Unexplained, Explained, and Total Deviation

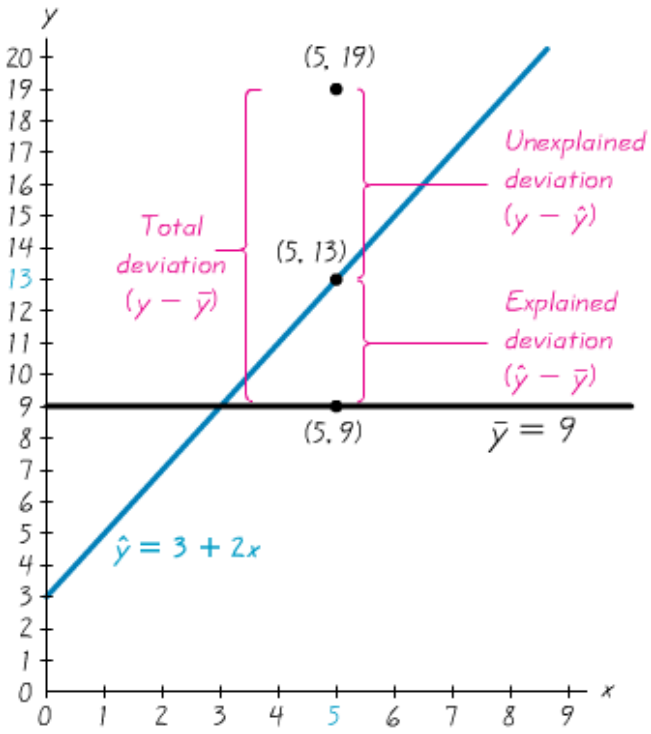
Definitions

Assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y-values is $\bar{y}$.

The ***total deviation*** of (x, y) is the vertical distance $y - \bar{y}$, which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$.

The ***explained deviation*** is the vertical distance $\hat{y} - \bar{y}$, which is the distance between the predicted y-value and the horizontal line passing through the sample mean $\bar{y}$.

The ***unexplained deviation*** is the vertical distance $y - \hat{y}$, which is the vertical distance between the point (x, y) and the regression line. (The distance $y - \hat{y}$ is also called a ***residual***.



- ✓ The mean of the y-value is given by $\bar{y} = 9$
- ✓ One of the pairs of sample data is $x = 5$ *and* $y = 19$
- ✓ The point (5, 13) is one of the points on the regression line, because substituting $x = 5$ into the regression equation of $\hat{y} = 3 + 2x$ yields $\hat{y} = 3 + 2(5) = 13$

Formula

$$(\textit{total variation}) = (\textit{explained variation}) + (\textit{unexplained variation})$$

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

Definition

The coefficient of determination is the amount of the variation in *y* that is explained by the regression line.

$$r^2 = \frac{\textit{explained variation}}{\textit{total variation}}$$

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Example

We used the paired pizza/subway fare costs to get $r = 0.988$. Find the coefficient of determinant. Also, find the percentage of the total variation in y (subway fare) that can be explained by the linear relationship between the cost of a slice pizza and the cost of a subway fare.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

The coefficient of determinant is $r^2 = 0.988^2 = 0.976$

Because r^2 is the proportion of total variation that is explained, we conclude that 97.6% of the total variation in subway fares can be explained by the cost of a slice of pizza. This means that 2.4% of the total variation in cost of subway fares can be explained by factors other than the cost of a slice of pizza.

Definition

A ***prediction interval***, is an interval estimate of a predicted value of *y*.

Definition

The ***standard error of estimate***, denoted by s_e is a measure of the differences (or distances) between the observed sample *y*-values and the predicted values $\hat{y}$ that are obtained using the regression equation.

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} \qquad (\textit{where } \hat{y} \textit{ is the predicted y-value})$$

Or

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

Example

Find the standard error of estimate s_e for the paired pizza/subway fare data listed below.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

x	y	xy	x^2	y^2
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1	1	1	1	1
1.25	1.35	1.6875	1.5625	1.8225
1.75	1.5	2.625	3.0625	2.25
2	2	4	4	4
6.5	6.35	9.4575	9.77	9.2175

$n = 6, \quad b_0 = 0.034560171, \quad b_1 = 0.94502138$

	Coefficients
Intercept	0.0345602
X Variable	0.9450214

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

$$= \sqrt{\frac{9.2175 - (0.034560171)(6.35) - (0.94502138)(9.4575)}{6-2}}$$

$$\approx 0.123$$

Prediction Interval for an Individual *y*

Given the fixed value x_0 the prediction interval for an individual *y* is

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

Where the margin of error *E* is

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

And x_0 represents the given value of *x*

$t_{\alpha/2}$ has $n - 2$ degrees of freedom

s_e is given in the previous formula

Example

For the paired pizza/subway fare costs from the Chapter Problem, we have found that for a pizza cost of $2.25, the best predicted cost of a subway fare is $2.16. Construct a 95% prediction interval for the cost of a subway fare, given that a slice of pizza costs $2.25 (so that $x = 2.25$).

Solution

$s_e = 0.122987$

$n = 6, \quad \bar{x} = \frac{6.5}{6} = 1.083333 \quad \sum x = 6.5 \quad \sum x^2 = 9.77$

$\alpha = 0.05$ (2-tails)

$t_{\alpha/2} = 2.776 \quad df = 6 - 2 = 4$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n\left(x_0 - \bar{x}\right)^2}{n\left(\sum x^2\right) - \left(\sum x\right)^2}}$$

$$= (2.776)(0.122987)\sqrt{1 + \frac{1}{6} + \frac{6(2.25 - 1.0833333)^2}{6(9.77) - (6.5)^2}}$$

≈ 0.441

With $\hat{y} = 2.16$ *and* $E = 0.441$

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

$2.16 - 0.441 < y < 2.16 + 0.441$

$1.72 < y < 2.60$

If the cost of a slice of pizza is $2.25, we have 95% confidence that the cost of a subway fare is between $1.72 and $2.60. That is a fairly large range of possible values, and one major factor contributing to the large range is that the sample size is very small with $n = 6$.

TI-83/84 PLUS The TI-83/84 Plus calculator can be used to find the linear correlation coefficient r, the equation of the regression line, the standard error of estimate s_e, and the coefficient of determination (labeled r^2). Enter the paired data in lists L1 and L2, then press **STAT** and select **TESTS,** and then choose the option **LinRegTTest.** For Xlist enter L1, for Ylist enter L2, use a Freq (frequency) value of 1, and select $\neq 0$. Scroll down to Calculate, then press the **ENTER** key.

Exercises *Section* 4.7 – Confidence and Prediction Intervals

1. A height of 70 in. is used to find the predicted weight is 180 lb. In your own words, describe a prediction interval in this situation.

2. A height of 70 in. is used to find the predicted weight is 180 lb. What is the major advantage of using a prediction interval instead of the predicted weight of 180 lb.? Why is the terminology of prediction interval used instead of confidence interval?

3. Use the value of the linear correlation $r = 0.873$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = tar\ in\ menthol\ cigarettes$

$y = movie\ gross$														
16	13	16	9	14	13	12		.14	13	13	16	13	13	18
9	19	2	13	14	14	15	16	6.						

1.1	0.8	1	0.9	0.8	0.8	0.8	0.8	0.9	0.8	0.8	1.2	0.8	0.8	1.3
0.7	1.4	0.2	0.8	1	0.8	0.8	1.2	0.6	0.7					

4. Use the value of the linear correlation

5. $r = 0.744$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = movie\ budget$

41	20	116	70	75	52	120	65	6.5	60	125	20	5	150
4.5	7	100	30	225	70	80	40	70	50	74	200	113	68
72	160	68	29	132	40								

117	5	103	66	121	116	101	100	55	104	213	34	12	290
47	10	111	100	322	19	117	48	228	47	17	373	380	118
114	120	101	120	234	209								

6. Use the value of the linear correlation $r = -0.865$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = car\ weight, \quad y = city\ fuel\ consumption\ in\ mi/gal$

7. Use the value of the linear correlation $r = -0.488$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = age\ of\ home, \quad y = home\ selling\ price$

8. Refer to the display obtained by using the paired data consisting of weights (in *lb*.) of 32 cars and their highway fuel consumption amounts (in *mi/gal*). A car weight of 4000 *lb*. to be used for predicting the highway fuel consumption amount

```
The regression equation is
Highway = 50.5 - 0.00587 Weight

Predictor         Coef     SE Coef      T      P
Constant        50.502       2.860  17.66  0.000
Weight     -0.0058685   0.0007859  -7.47  0.000

S = 2.19498   R-Sq = 65.0%   R-Sq(adj) = 63.9%

Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  27.028   0.497  (26.013, 28.042)  (22.431, 31.624)

Values of Predictors for New Observations
New
Obs  Weight
  1    4000
```

a) What percentage of the total variation in highway fuel consumption can be explained by the linear correlation between weight and highway fuel consumption?

b) If a car weighs 4000 *lb*., what is the single value that is the best predicted amount of highway fuel consumption? (Assume that there is a linear correlation between weight and highway fuel consumption.)

9. The paired values of the Consumer Price Index (CPI) and the cost of a slice of pizza are shown below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a) Find the explained variation

b) Find the unexplained variation

c) Find the total variation

d) Find the coefficient of determination

e) Find the standard error of estimate s_e

f) Find the predicted cost of a slice of pizza for the year 2001, when the CPI was 187.1.

g) Find a 95% prediction interval estimate of the cost of a slice of pizza when the CPI was 187.1

In each case, there is sufficient evidence to support a claim of a linear correlation so that it is reasonable to use the regression equation when making predictions.

10. Find the best predicted temperature for a recent year in which the concentration (in parts per million) of CO_2 and temperature (in °C) for different years

CO_2	314	317	320	326	331	339	346	354	361	369
Temperature	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

a) Find the explained variation

b) Find the unexplained variation

c) Find the total variation

d) Find the coefficient of determination

e) Find the standard error of estimate s_e

f) Find the predicted temperature (in °C) when CO_2 concentration is 370.9 parts per million.

g) Find a 99% prediction interval estimate temperature (in °C) when CO_2 concentration is 370.9 parts per million

In each case, there is sufficient evidence to support a claim of a linear correlation so that it is reasonable to use the regression equation when making predictions.

Find a prediction interval data listed below.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

11. Using: *Cost of a slice of pizza*: \$2.10; 99% *confidence*

12. Using: *Cost of a slice of pizza*: \$2.10; 90% *confidence*

13. Using: *Cost of a slice of pizza*: \$0.50; 95% *confidence*

14. Using: *Cost of a slice of pizza*: \$0.75; 99% *confidence*