Solution   Section 4.7 – Variation and Prediction Intervals

Exercise

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.

Solution

A prediction interval is an interval estimate for a predicted value. In this situation it will be a range of weights centered at the prediction's point estimate of 180 lbs.

Exercise

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?

Solution

By providing a range of values instead of a single point, a prediction interval gives an indication of the accuracy of the prediction. A confidence interval is an internal estimate of a parameter – i.e., of a conceptually fixed, although unknown, value. A prediction interval is an interval estimate of a random variable – i.e., of a value from a distribution of values.

Exercise

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

$x = tar\ in\ menthol\ cigarettes$

16	13	16	9	14	13	12	14	14	13	13	16	13	13	18
9	19	2	13	14	14	15	16	6	8					

$y = nicotine\ in\ menthol\ cigarettes$

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					

Solution

The coefficient of determination is $r = (0.873)^2 = 0.762$

The portion of the total variation in y explained by the regression is $r^2 = 0.762 \underline{= 76.2\%}$

Regression Statistics	
Multiple R	0.873034386
R Square	0.762189039
Adjusted R Square	0.751849432
Standard Error	0.120760017
Observations	25

Exercise

Use the value of the linear correlation $r = 0.744$ 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								

$y = movie\ gross$

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								

Solution

The coefficient of determination is $r = (0.744)^2 = 0.554$

The portion of the total variation in y explained by the regression is $r^2 = 0.554 \underline{= 55.4\%}$

Exercise

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

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

Solution

The coefficient of determination is $r = (-0.865)^2 = 0.748$

The portion of the total variation in y explained by the regression is $r^2 = 0.748 \underline{= 74.8\%}$

Exercise

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

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

Solution

The coefficient of determination is $r = (-0.488)^2 = 0.238$

The portion of the total variation in y explained by the regression is $r^2 = 0.238 \underline{= 23.8\%}$

Exercise

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.)

Solution

a) $\underline{R - squared = 65.0\%}$

b) The given point estimate is $\underline{\hat{y} = 27.028\ \ mpg}$

Exercise

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.

Solution

The predicted values:

	Coefficients
Intercept	-0.161601
X Variable	0.0100574

$\hat{y} = -0.161601 + 0.0100574x$

x	**y**	$\hat{y}$	$\bar{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$	$y-\hat{y}$	$(y-\hat{y})^2$	$y-\bar{y}$	$(y-\bar{y})^2$
30.2	0.15	0.142	1.083	–0.940	0.886	0.008	0.000	–0.930	0.871
48.3	0.35	0.324	1.083	–0.760	0.576	0.023	0.001	–0.730	0.538
112.3	1.00	0.968	1.083	–0.120	0.013	0.032	0.001	–0.080	0.007
162.2	1.25	1.470	1.083	0.386	0.149	–0.220	0.048	0.167	0.028
191.9	1.75	1.768	1.083	0.685	0.469	–0.018	0.000	0.667	0.444
197.8	2.00	1.828	1.083	0.744	0.554	0.172	0.030	0.917	0.840
742.7	6.50	6.50	6.50	0.0	2.648	0.0	0.08	0.0	2.728

a) The explained variation is $\sum(\hat{y}-\bar{y})^2 = 2.648$

b) The unexplained variation is $\sum(y-\hat{y})^2 = 0.080$

c) The total variation is $\sum(y-\bar{y})^2 = 2.728$

d) $r^2 = \dfrac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \dfrac{2.648}{2.728} = 0.971$

e) $s_e^2 = \frac{\sum (y-\hat{y})^2}{n-2} = \frac{0.08}{4} = 0.02$

$s_e = \sqrt{0.02} = 0.141$

f) $\hat{y}\big|_{187.1} = -0.161601 + 0.0100574(187.1)$

$= 1.7201$

$= \$1.72$

g) Preliminary calculations for $n = 6$

x	y	x^2
30.2	0.15	912.04
48.3	0.35	2332.89
112.3	1.00	12611.29
162.2	1.25	26308.84
191.9	1.75	36825.61
197.8	2.00	39124.84
742.7	6.50	118115.5

$\bar{x} = \frac{\sum x}{n} = \frac{742.7}{6} = 123.783$

$\alpha = 0.05$ (2-tails)

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

TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail 0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails 0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	$\boxed{2.776}$	2.132	1.533

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

$$= (2.776)(0.141)\sqrt{1 + \frac{1}{6} + \frac{6(187.1 - 123.783)^2}{6(118115.5) - (742.7)^2}}$$

$$\approx 0.4450$$

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

$1.7201 - 0.445 < y_{187.1} < 1.7201 + 0.445$

$\$1.27 < y_{187.1} < \$2.17$

Exercise

The paired values of the Consumer Price Index (CPI) and the subway fare are shown below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Subway fare	0.15	0.35	1.00	1.35	1.5	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 subway fare for the year 2001, when the CPI was 187.1.

g) Find a 95% prediction interval estimate of the cost of subway fare 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.

Solution

The predicted values (from Excel):

	Coefficients
Intercept	-0.124252712
X Variable 1	0.009553677

$$\hat{y} = -0.124253 + 0.00955368x$$

x	y	$\hat{y}$	$\bar{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$	$y-\hat{y}$	$(y-\hat{y})^2$	$y-\bar{y}$	$(y-\bar{y})^2$
30.2	0.15	0.164	1.058	–0.890	0.799	–0.014	0.0	–0.910	0.825
48.3	0.35	0.337	1.058	–0.720	0.520	0.013	0.0	–0.710	0.502
112.3	1.00	0.949	1.058	–0.110	0.012	0.051	0.006	0.292	0.085
162.2	1.35	1.425	1.058	0.367	0.135	–0.075	0.006	0.292	0.085
191.9	1.50	1.709	1.058	0.651	0.423	–0.209	0.044	0.442	0.195
197.8	2.00	1.765	1.58	0.707	0.500	0.235	0.055	0.942	0.887
742.7	6.35	6.350	6.350	0.0	2.930	0.0	0.104	0.0	2.497

a) The explained variation is $\sum(\hat{y}-\bar{y})^2 = \underline{2.390}$

b) The unexplained variation is $\sum(y-\hat{y})^2 = \underline{0.107}$

c) The total variation is $\sum(y-\bar{y})^2 = \underline{2.497}$

d) $r^2 = \dfrac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \dfrac{2.648}{2.728} \underline{= 0.957}$

Regression Statistics	
Multiple R	0.978255696
R Square	0.956984207
Adjusted R Square	0.946230258
Standard Error	0.163870391
Observations	6

e) $s_e^2 = \dfrac{\sum(y-\hat{y})^2}{n-2} = \dfrac{0.107}{4} = 0.02675$

$s_e = \sqrt{0.02675} = 0.164$

f) $\hat{y}\big|_{187.1} = -0.124253 + 0.00955368(187.1)$

$= \$1.66$

g) Preliminary calculations for $n = 6$

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{742.7}{6} = 123.783$

$\alpha = 0.05 \quad \text{(2-tails)}$

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

x	y	x^2
30.2	0.15	912.04
48.3	0.35	2332.89
112.3	1.00	12611.29
162.2	1.35	26308.84
191.9	1.50	36825.61
197.8	2.00	39124.84
742.7	6.35	118115.51

TABLE A-3 t Distribution: Critical t Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	**Area in Two Tails**				
	0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	2.776	2.132	1.533

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

$$= (2.776)(0.164)\sqrt{1 + \frac{1}{6} + \frac{6(187.1 - 123.783)^2}{6(118115.5) - (742.7)^2}}$$

$$\approx 0.5230$$

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

$1.6632 - 0.5230 < y_{187.1} < 1.6632 + 0.5230$

$\$1.14 < y_{187.1} < \$2.19$

Exercise

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.

Solution

The predicted values (from Excel):

	Coefficients
Intercept	10.48308065
X Variable 1	0.010917736

$\hat{y} = 10.4831 + 0.0109177x$

x	**y**	$\hat{y}$	$\bar{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$	$y-\hat{y}$	$(y-\hat{y})^2$	$y-\bar{y}$	$(y-\bar{y})^2$
314	13.9	13.911	14.17	–0.259	0.067	–0.011	0.0	–0.27	0.073
317	14	13.944	14.17	–0.266	0.051	0.056	0.003	–0.17	0.029
320	13.9	13.977	14.17	–0.193	0.037	–0.077	0.006	–0.27	0.073
326	14.1	14.042	14.17	–0.128	0.016	0.058	0.003	–0.07	0.005
331	14	14.097	14.17	–0.073	0.005	–0.097	0.009	–0.17	0.029
339	14.3	14.184	14.17	0.014	0.0	0.116	0.013	0.13	0.017
346	14.1	14.261	14.17	0.091	0.008	–0.161	0.026	–0.07	0.005
354	14.5	14.348	14.17	0.178	0.032	0.152	0.023	0.33	0.109
361	14.5	14.424	14.17	0.254	0.065	0.076	0.006	0.33	0.109
369	14.4	14.512	14.17	0.342	0.117	–0.112	0.012	0.23	0.053
3377	141.7	141.7	141.70	0.0	0.399	0.0	0.102	0.0	0.501

a) The explained variation is $\sum(\hat{y}-\bar{y})^2 \underline{= 0.399}$

b) The unexplained variation is $\sum(y-\hat{y})^2 \underline{= 0.102}$

c) The total variation is $\sum (y-\bar{y})^2 = 0.501$

d) $$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{0.399}{0.501} = 0.796$$

e) $$s_e^2 = \frac{\sum(y-\hat{y})^2}{n-2} = \frac{0.102}{8} = 0.01275$$

$$s_e = \sqrt{0.01275} = 0.113$$

f) $$\hat{y}\big|_{370.9} = 10.4831 + 0.0109177(370.9)$$

$$= 14.53\ ^\circ C$$

g) Preliminary calculations for $n = 8$

$$\bar{x} = \frac{\sum x}{n} = \frac{3377}{10} = 337.7$$

$$\alpha = 0.01 \quad and \quad df = n-2 = 8 \quad \text{(2-tails)}$$

$$t_{\alpha/2} = t_{0.005} = 3.355$$

Regression Statistics	
Multiple R	0.891976355
R Square	0.795621818
Adjusted R Square	0.770074545
Standard Error	0.113133477
Observations	10

x	y	x^2
314	13.9	985696
317	14	100489
320	13.9	102400
326	14.1	106276
331	14	109561
339	14.3	114921
346	14.1	119716
354	14.5	125316
361	14.5	130321
369	14.4	136161
3377	141.7	1143757

TABLE A-3 t Distribution: Critical t Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
8	3.355	2.896	2.306	1.860	1.397

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

$$= (3.355)(0.113)\sqrt{1 + \frac{1}{10} + \frac{10(370.9-337.7)^2}{10(1143757) - (3377)^2}}$$

$$\approx 0.4533$$

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

$$14.5324 - 0.4533 < y_{370.9} < 14.5324 + 0.4533$$

$$14.08\ ^\circ C < y_{370.9} < 14.99\ ^\circ C$$

Exercise

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

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

Solution

The predicted values (from Excel):

	Coefficients
Intercept	0.03456017
X Variable 1	0.94502138

$$\hat{y} = 0.034560 + 0.945021x$$

$$\hat{y}\Big|_{2.1} = 0.034560 + 0.945021(2.1)$$

$$= 2.019$$

$\alpha = 0.01 \quad and \quad df = n - 2 = 4$

$t_{\alpha/2} = t_{0.005} = 4.604$

TABLE A-3	*t* Distribution: Critical *t* Values				
	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	2.776	2.132	1.533

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

$$= (4.604)(0.122987)\sqrt{1 + \frac{1}{6} + \frac{6(2.1 - 1.083333)^2}{6(9.77) - (6.5)^2}}$$

$$\approx 0.704$$

x	x^2
0.15	0.0225
0.35	0.1225
1	1
1.25	1.5625
1.75	3.0625
2	4
6.5	9.77

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

$$2.019 - 0.704 < y_{2.1} < 2.019 + 0.704$$

$$\$1.32 < y_{2.1} < \$2.72$$

Exercise

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

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

Solution

The predicted values (from Excel):

	Coefficients
Intercept	0.03456017
X Variable 1	0.94502138

$$\hat{y} = 0.034560 + 0.945021x$$

$$\hat{y}\Big|_{2.1} = 0.034560 + 0.945021(2.1)$$

$$= 2.019$$

Regression Statistics	
Multiple R	0.98781094
R Square	0.97577045
Adjusted R Square	0.96971306
Standard Error	0.122987
Observations	6

$$\alpha = 0.1 \quad \textit{and} \quad df = n - 2 = 4$$

$$t_{\alpha/2} = t_{0.05} = 2.132$$

TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	2.776	2.132	1.533

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

$$= (2.132)(0.122987)\sqrt{1 + \frac{1}{6} + \frac{6(2.10 - 1.083333)^2}{6(9.77) - (6.5)^2}}$$

$$\approx 0.326$$

x	x^2
0.15	0.0225
0.35	0.1225
1	1
1.25	1.5625
1.75	3.0625
2	4
6.5	9.77

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

$$2.019 - 0.326 < y_{2.1} < 2.019 + 0.326$$

$$\$1.69 < y_{2.1} < \$2.34$$

Exercise

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

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

Solution

The predicted values (from Excel):

	Coefficients
Intercept	0.03456017
X Variable 1	0.94502138

$$\hat{y} = 0.034560 + 0.945021x$$

Regression Statistics	
Multiple R	0.98781094
R Square	0.97577045
Adjusted R Square	0.96971306
Standard Error	0.122987
Observations	6

$$\hat{y}\Big|_{0.50} = 0.034560 + 0.945021(0.5)$$

$$= 0.507$$

$\alpha = 0.05 \quad and \quad df = n - 2 = 4$

$t_{\alpha/2} = t_{0.025} = 2.776$

TABLE A-3	*t* Distribution: Critical *t* Values				
			Area in One Tail		
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom			Area in Two Tails		
	0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	2.776	2.132	1.533

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

$$= (2.776)(0.122987)\sqrt{1 + \frac{1}{6} + \frac{6(0.5 - 1.083333)^2}{6(9.77) - (6.5)^2}}$$

$$\approx 0.388$$

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

$$0.507 - 0.388 < y_{0.5} < 0.507 + 0.388$$

$$\$0.12 < y_{0.5} < \$0.89$$

Exercise

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

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

Solution

The predicted values (from Excel):

	Coefficients
Intercept	0.03456017
X Variable 1	0.94502138

$$\hat{y}=0.034560+0.945021x$$

$$\hat{y}\big|_{0.75}=0.034560+0.945021(0.75)$$

$$=0.743$$

$\alpha=0.01 \quad and \quad df=n-2=4$

$t_{\alpha/2}=t_{0.005}=4.604$

Table A-3	*t* Distribution: Critical *t* Values				
	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
4	**4.604**	3.747	**2.776**	2.132	1.533

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

$$=(4.604)(0.122987)\sqrt{1+\frac{1}{6}+\frac{6(0.75-1.083333)^2}{6(9.77)-(6.5)^2}}$$

$$\approx 0.622$$

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

$$0.743-0.622<y_{0.75}<0.743+0.622$$

$$\$0.12<y_{0.75}<\$1.37$$