

4.2 Linear Regression and the Coefficient of Determination with work

HW: pg. 154: 1 - 7(o), 13, 15

1. Explanatory variables is placed along the horizontal axis. Response variables is placed along the vertical axis.
 3. Decreases.
 5. a) Moderate b) None c) High
 7. a) None b) Increasing population might be a lurking variable causing both variables to increase.
 13. a) Lines slopes upward. b) Strong, positive c) $r = 0.972$; increase.

L1	L2	L3	Z
3	88		
4	95		
5	110		
6	125		
7	140		
8	155		
9	170		
10	185		
11	200		
12	215		
13	230		
14	245		
15	260		
16	275		
17	290		
18	305		
19	320		
20	335		
21	350		
22	365		
23	380		
24	395		
25	410		
26	425		
27	440		
28	455		
29	470		
30	485		
31	500		
32	515		
33	530		
34	545		
35	560		
36	575		
37	590		
38	605		
39	620		
40	635		
41	650		
42	665		
43	680		
44	695		
45	710		
46	725		
47	740		
48	755		
49	770		
50	785		
51	800		
52	815		
53	830		
54	845		
55	860		
56	875		
57	890		
58	905		
59	920		
60	935		
61	950		
62	965		
63	980		
64	995		
65	1010		
66	1025		
67	1040		
68	1055		
69	1070		
70	1085		
71	1100		
72	1115		
73	1130		
74	1145		
75	1160		
76	1175		
77	1190		
78	1205		
79	1220		
80	1235		
81	1250		
82	1265		
83	1280		
84	1295		
85	1310		
86	1325		
87	1340		
88	1355		
89	1370		
90	1385		
91	1400		
92	1415		
93	1430		
94	1445		
95	1460		
96	1475		
97	1490		
98	1505		
99	1520		
100	1535		

$y = a + bx$
 $a = 55.73170732$
 $b = 0.894308943$
 $r^2 = 0.94522448$
 $r = 0.9722460841$

15. a) line slopes downward b) strong, negative c) $r = -0.990$; decrease.

L1	L2	L3	Z
1000	90		
975	100		
950	85		
925	95		
900	80		
875	90		
850	100		
825	85		
800	95		
775	80		
750	90		
725	100		
700	85		
675	95		
650	80		
625	90		
600	100		
575	85		
550	95		
525	80		
500	90		
475	100		
450	85		
425	95		
400	80		
375	90		
350	100		
325	85		
300	95		
275	80		
250	90		
225	100		
200	85		
175	95		
150	80		
125	90		
100	100		
75	85		
50	95		
25	80		
0	90		
-25	100		
-50	85		
-75	95		
-100	80		
-125	90		
-150	100		
-175	85		
-200	95		
-225	80		
-250	90		
-275	100		
-300	85		
-325	95		
-350	80		
-375	90		
-400	100		
-425	85		
-450	95		
-475	80		
-500	90		
-525	100		
-550	85		
-575	95		
-600	80		
-625	90		
-650	100		
-675	85		
-700	95		
-725	80		
-750	90		
-775	100		
-800	85		
-825	95		
-850	80		
-875	90		
-900	100		
-925	85		
-950	95		
-975	80		
-1000	90		

$y = a + bx$
 $a = 1496.704604$
 $b = -1.442676906$
 $r^2 = 0.975891024$
 $r = -0.987373906$

Oct 16-9:36 AM

4.2 Linear Regression and the Coefficient of Determination

Essential Questions:

- How do we write an equation given data?
- How do I use the coefficient of determination to help understand data?

Focus Points:

- State the least-squares criterion.
- Use sample data to find the equation of the LSRL. Graph the LSRL.
- Use the least-squares line to predict a value of the response variable y for a specified value of the explanatory variable x .
- Explain the difference between interpolation and extrapolation.
- Explain why extrapolation beyond the sample data range might give results that are misleading or meaningless.
- Use r^2 to determine *explained* and *unexplained* variation of the response variable y .

Oct 17-3:15 PM

Least-Squares Criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.

Least squares criterion

The line of best fit minimises the total sum of squares of the vertical deviations for each case.

b = slope of the line of best fit

residuals = vertical (Y) distance between line of best fit and each observation

a = point at which line of best fit crosses the Y-axis. (unexplained variance)

Oct 17-3:24 PM

Least-Squares Regression Line (LSRL)

The "best-fitting" or trend line found using actual data. The line **minimizes** the sum of the squares of the **vertical distances** between the points and the line over **ALL** point in the scatter diagram. **Minimizes residuals.**

$$\hat{y} = a + bx$$

a = y-intercept
 b = slope

Oct 17-3:30 PM

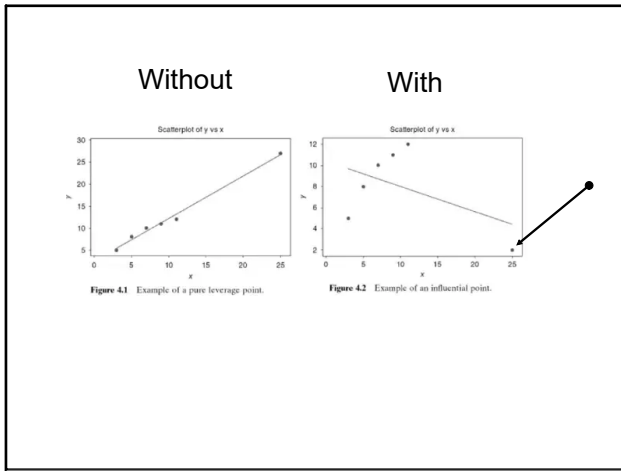
The **slope** of the least-squares line tells us how many units the response variable is expected to change for each unit change in the explanatory variable. *The number of units in the response variable for each unit change in the explanatory variable* is called the **marginal change** of the response variable.

Oct 18-7:33 AM

A data pair is **influential** if removing it would **substantially change** the equation of the least-squares line or other calculations associated with linear regression. An **influential point** often has an **x-value** near the **extreme high or low value** of the data set. Also known as an **outlier**.

Oct 18-7:35 AM

4.2 Linear Regression and the Coefficient of Determination with work



Oct 18-7:37 AM

USING THE LSRL FOR PREDICTION

This is the main point of regression. Using the \hat{y} for a specified x value, but the accuracy depends on many things.

- Are there any influential points?
- Consider the sample correlation coefficient "r". The closer r is to 1 or -1 the better the prediction.
- Consider the coefficient of determination r^2
- Look at the residuals and a residual plot

Oct 18-7:39 AM

The **residual** is the difference between the y -value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x .

$y - \hat{y}$ is the residual (put this in List 3 or List 4)

****The sum of the residuals should always equal zero!!****

Handwritten notes: L_1, L_2, L_3, L_4 above $x, y, \hat{y}, y - \hat{y}$; $L_2 - L_3$ below $y - \hat{y}$

Oct 18-7:47 AM

The **residual plot** uses the x values on the horizontal axis and the $y - \hat{y}$ (residuals) on the vertical axis.

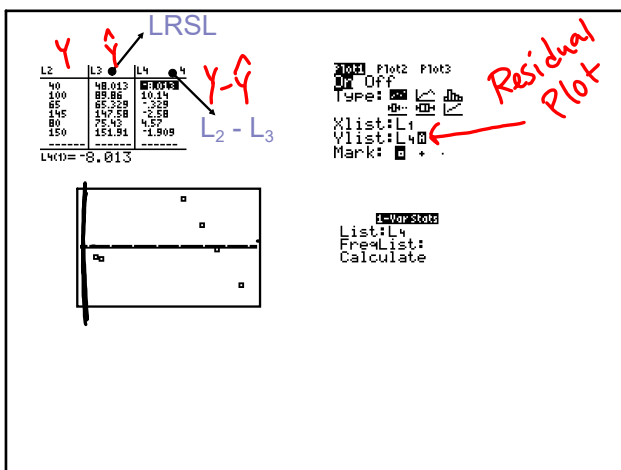
The mean of the residuals is **ALWAYS ZERO**.

$L_1 = x$ -values
 $L_2 = y$ -values
 $L_3 = \hat{y}$: LSRL
 $L_4 = y - \hat{y}$ (residuals)

Then graph a scatter diagram using L_1 and L_4

Handwritten: x and y with arrows pointing to the axes.

Oct 18-8:37 AM



Oct 18-8:42 AM

Predicting \hat{y} values for x values that are **between** observed x values in the data set is called **interpolation**.

Predicting \hat{y} values for x values that are **beyond** observed x values in the data set is called **extrapolation**. Extrapolation may produce unrealistic forecasts.

Oct 18-7:52 AM

4.2 Linear Regression and the Coefficient of Determination with work

COEFFICIENT OF DETERMINATION

r^2

1. Compute the sample correlation coefficient r using the calculator. Then square it to get the coefficient of determination.
2. The value of r^2 is the **ratio of explained variation over total variation**. That is, r^2 is the fractional amount of total variation in y that can be explained by using the LSRL.
3. Furthermore, $1 - r^2$ is the fractional amount of total variation in y that is **unexplained variation due to random chance** or due to the possibility of **lurking variable** that influence y .
4. If $r^2 = .81$, then we can say that 81% of the variation/behavior of the y variable **can be explained** and 19% **cannot** be explained.

Oct 18-8:22 AM

Example: Car Dealership

The Quick Sell car dealership has been using 1-minute spot ads on a local TV station. The ads always occur during the evening hours and advertise the different models and price ranges of cars on the lot that week. During a 10-week period, a Quick Sell dealer kept a weekly record of the number x of TV ads versus the number y of cars sold.

Week	TV Ads (x)	Cars Sold (y)
1	6	15
2	20	31
3	10	10
4	16	28
5	28	20
6	40	40
7	25	12
8	18	15
9	10	8
10	8	15

- Calculate the LSRL. $\hat{y} = 6.541 + 1.011x$
- Graph a scatter plot with your LSRL.
- The manager decided that Quick Sell can afford only 12 ads per week. At that level of advertisement, how many cars can Quick Sell expect to sell each week? $x=12; \hat{y}=?; f(12) = 6.541 + 1.011(12) = 18.673$
- How reliable is your answer from part c? Is it interpolation or extrapolation? $r = .919$; so very reliable
- What is your coefficient of determination? What does this value tell us? $r^2 = .845$ 85% explained & 15% not explained

Oct 18-8:08 AM

CALCULATOR TIPS!!

Calculator screen showing regression statistics and a scatter plot with a line of best fit. The screen displays various regression options like 1:1-Var Stats, 2:2-Var Stats, 3:Med-Med, 4:LinReg(ax+b), 5:QuadReg, 6:CubicReg, 7:QuartReg, 8:LinReg(a+bx), and 9:Med-Med. It also shows regression statistics such as \bar{x} , \bar{y} , s_x , s_y , r , and r^2 .

r & r^2 : 2nd 0: Diagnostics ON

Oct 18-8:32 AM

HW: pg. 171: 1, 3, 7, 9, 11, 19

- b = -2. When x changes by 1 unit, y decreases by 2 units.
- Extrapolating. Extrapolating beyond the range of the data is dangerous because the relationship pattern might change.
- a & d) b) calc c) $\bar{x} \approx 33.67$ jobs, $\bar{y} \approx 4.67$ entry-level jobs
 $a \approx -0.748$, $b \approx 0.161$
 $\hat{y} = -0.748 + 0.161x$
 e) $r^2 = 0.740$, 74% explained 26% unexplained
 f) 5.69 jobs
- Weight of Cars and Gasoline Mileage b) calc c) $\bar{x} \approx 37.375$, $\bar{y} \approx 20.875$ mpg
 $a \approx 43.326$, $b \approx -0.601$
 $\hat{y} = 43.326 - 0.601x$
 e) $r^2 \approx 0.895$, 89.5% explained 10.5% unexplained
 f) 20.5 mpg
- Age and Percentage of Total Population That is Working b) calc c) $\bar{x} \approx 47$ years, $\bar{y} \approx 16.43\%$
 $a \approx 39.761$, $b \approx -0.496$
 $\hat{y} = 39.761 - 0.496x$
 e) $r^2 \approx 0.920$, 92% explained 8% unexplained
 f) 27.36%
- a) Yes. The pattern of residuals appears randomly scattered about the horizontal line at 0.
 b) No. There do not appear to be any outliers.

Oct 18-8:30 AM