4.2 Linear Regression and the Coefficient of Determination with work

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

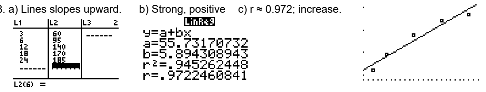
1. Explanatory variables is placed along the horizontal axis. Response variables is placed along the vertical axis.

3. Decreases.

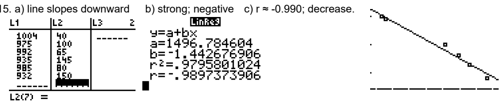
5. a) Moderate   b) None   c) High

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.



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



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?

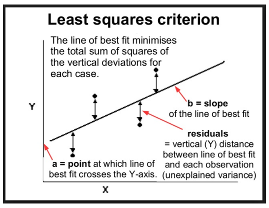
Focus Points:

- 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.



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

Without   With

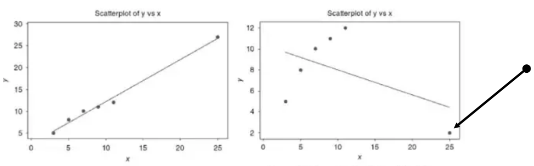


Figure 4.1 Example of a pure leverage point.   Figure 4.2 Example of an influential point.

Oct 18-7:37 AM

USING THE LSRL FOR PREDICTION

This is the main point of regression. Using the y-hat 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

L_1	L_2	L_3	L_4
x	y	$\hat{y}$	$y-\hat{y}$: $L_2 - L_3$

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 is List 3 or List 4)

The sum of the residuals should always equal zero!!

Oct 18-7:47 AM

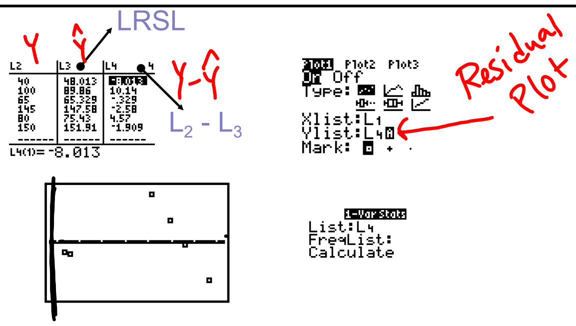
The residual plot uses the x values on the horizontal axis and the $y - \hat{y}$ (residuals) on the vertical axis.

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
↓ x   ↓ y

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

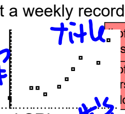
1. Compute the sample correlation coefficient r using the calculator. Then square it to get the coefficient of determination.
2. The value of r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of total variation in y that can be explained by using the 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.

of TV s	6	20	0	14	25	16	28	18	10	8
of rs ld	15	31	10	16	28	20	40	25	12	15



a) Calculate the LRSL.

$\hat{y} = 6.541 + 1.011x$

b) Graph a scatter plot with your LSRL.

c) 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?

$x = 12;\ \hat{y} = ?$ $\quad f(12) = 6.541 + 1.011(12) = 18.673$ $\boxed{\approx 18 \text{ cars}}$

d) How reliable is your answer from part c? Is it interpolation or extrapolation?

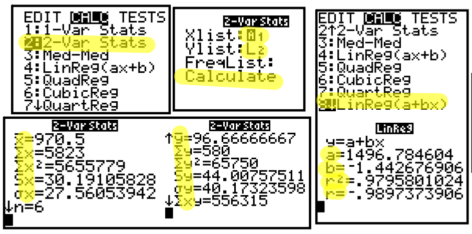
$r = .919$; so very reliable

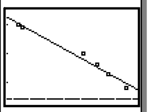
e) What is your coefficient of determination? What does this value tell us?

$r^2 = .845$ 85% explained & 15% not explained

Oct 18-8:08 AM

CALCULATOR TIPS!!





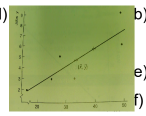
r & r^2 : 2nd 0: Diagnostics ON

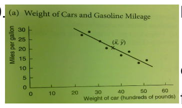
Oct 18-8:32 AM

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

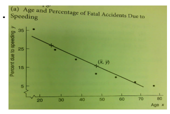
1. b = -2. When x changes by 1 unit, y decreases by 2 units.

3. Extrapolating. Extrapolating beyond the range of the data is dangerous because the relationship pattern might change.

7. a & d) 
b) calc c) $\bar{x} \approx 33.67$ jobs, $\bar{y} \approx 4.67$ entry-level jobs
a ≈ -0.748, b ≈ 0.161,
$\hat{y}$ = -0.748 + 0.161x
e) r^2 = 0.740, 74% explained 26% unexplained
f) 5.69 jobs

9. 

b) calc c) $\bar{x} \approx 37.375$, $\bar{y} \approx 20.875 mpg$
a ≈ 43.326, b ≈ -0.601,
$\hat{y}$ ≈ 43.326-0.601x
e) r^2 ≈ 0.895, 89.5% explained 10.5% unexplained
f) 20.5 mpg

11. 
b) calc c) $\bar{x} \approx 47 years$, $\bar{y} \approx 16.43\%$
a ≈ 39.761, b ≈ -0.496,
$\hat{y}$ ≈ 39.761- 0.496x
e) r^2 ≈ 0.920, 92% explained 8% unexplained
f) 27.36%

19. 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