AP Statistics Notes – Unit Three: Exploring Relationships Between Variables

Syllabus Objectives: 1.12 – The student will analyze patterns in scatterplots. 1.13 – The student will assess the linearity of bivariate data.

Scatterplots

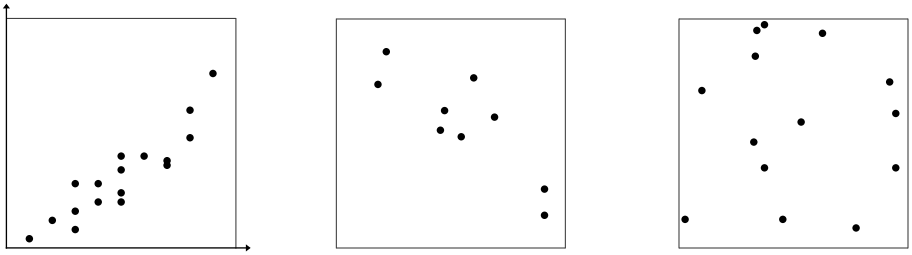
- This is the most effective way to display the relationship between two QUANTITATIVE variables.
- It shows the RELATIONSHIP between two quantitative variables (bivariate data) measured on the same individuals.
- Determine the explanatory and response variables.
    - A **response variable** measures an outcome of a study. The response variable is also called the dependent variable and is graphed on the vertical (*y*) axis.
    - An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is called the independent variable and is graphed on the horizontal (*x*) axis.
- If there is no explanatory-response distinction between the variables, either variable can go on either axis.
- Plot the individual order pairs.
- Interpret the scatterplot.
- **Ex:** A sample of one-way Greyhound bus fares from Rochester NY to cities less than 750 miles was taken from the Greyhound website. The following table gives the destination city, the distance and the one-way fare. Distance is the explanatory variable and Fare is the response variable. Fare should *depend* on distance.

Destination City	Distance	Standard One-Way Fare
Albany, NY	240	39
Baltimore, MD	430	81
Buffalo, NY	69	17
Chicago, IL	607	96
Cleveland, OH	257	61
Montreal, QU	480	70.5
New York City, NY	340	65
Ottawa, ON	467	82
Philadelphia, PA	335	67
Potsdam, NY	239	47
Syracuse, NY	95	20
Toronto, ON	178	35
Washington, DC	496	87

***Comments: The axes need not intersect at (0,0). For each of the axes, the scale should be chosen so that the minimum and maximum values on the scare are convenient and the values to be plotted are between the two values. For this example, the x axis (distance) runs from 50 to 650 miles where the data points are between 69 and 607. The y axis (fare) runs from $10 to $100 where the data points are between $17 and $96. Each axis should also be appropriate labeled.

Interpreting Scatterplots

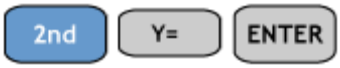
- Look for the **overall pattern** and for any striking **deviations** from that pattern.
  - An important kind of deviation is an **outlier**, an individual ordered pair that falls outside the overall pattern on the relationship.
- Describe the overall pattern by the **form, direction** and **strength** of the relationship.
  - **Form –** is it linear or curved?
  - **Direction –** does the overall pattern moves from upper left to lower right, or vice versa?
    - Two variables are **positively associated** when as one variable increases, the other increases as well. The pattern moves from lower left to upper right.
    - Two variables are **negatively associated** if one variable increases and the other decreases. The pattern moves from upper left to lower right.
    -



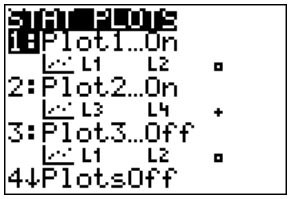
Positive Association Negative Association No Association

  - **Strength –** How closely do the points follow a clear form? Linear strength is normally described as weak, moderate or strong. The more linear, the stronger it is.

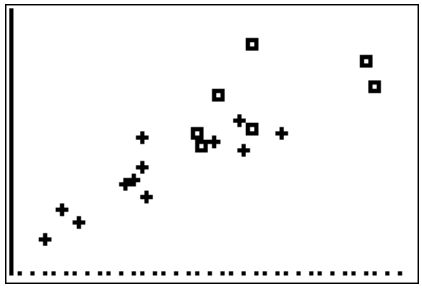
- Categorical variables can be added to scatterplots. Use different colors or symbols to plot points when adding a categorical variable (to see a pattern in males or females separately – plot males as a dot and females as a square.)
  - **Ex:** Do heavier people burn more energy? Metabolic rate, the rat at which the body consumes energy, is important in studies of weight gain, dieting and exercise. Data on the lean body mass and resting metabolic rate for 12 women and 7 men who were subjects in a study of dieting were input in the TI-84 graphing calculator. The lean body mass for males was placed in L_1, the male metabolic rate was placed in L_2. The mass and metabolic rate for females were placed in L_3 and L_4. Press



and turn two of the plots on.



The first plot will graph the male data as "squares" and the second plot will graph the female data as "plus signs". Hit Zoom 9 or set the window to display the scatterplot.



Rate (cal)

Body Mass (kg)

Syllabus Objective: 1.14 – The student will calculate the coefficients of correlation and determination.

Measuring Linear Association: Correlation

- **Correlation** measures the strength and direction of the linear association between two quantitative variables.
  - The variable that represents correlation is r.
  - r measures how tightly the points on a scatterplot cluster about a straight line.
  - r is called the **correlation coefficient.**
  - **Formula:**

$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum\left[\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right]}{n-1}$$

**Notice that the formula begins by standardizing the observations. r is the average of the products of the two standardized formulas. Therefore, r does not have any units.

  - **Ex:**

x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
240	39	-0.5214	-0.7856	0.4096
430	81	0.6357	0.8610	0.5473
69	17	-1.5627	-1.6481	2.5755
607	96	1.7135	1.4491	2.4831
257	61	-0.4178	0.0769	-0.0321
480	70.5	0.9402	0.4494	0.4225
340	65	0.0876	0.2337	0.0205
467	82	0.8610	0.9002	0.7751
335	67	0.0571	0.3121	0.0178
239	47	-0.5275	-0.4720	0.2489
95	20	-1.4044	-1.5305	2.1494
178	35	-0.8989	-0.9424	0.8472
496	87	1.0376	1.0962	1.1374
				11.6021

$\bar{x} = 325.615$

$s_x = 164.2125$

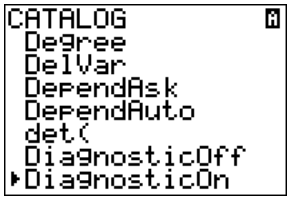
$\bar{y} = 59.0385$

$s_y = 25.506$

$$r = \frac{11.601}{13-1} = 0.9668$$

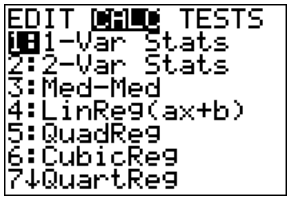
- Finding the correlation coefficient on the TI-84 graphing calculator:
  - Enter the explanatory variable in L_1 and enter the response variable in L_2.
  - Make sure your Diagnostics are on. This is NOT the calculator's default.
  - Press 2nd 0

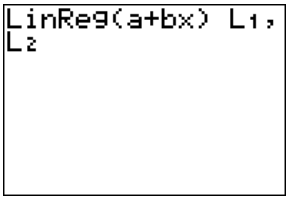
and scroll down until you find DiagnosticOn.



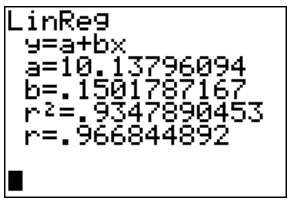
Hit enter twice. Now the Diagnostics are on.

  - Go to the Stat menu. Entry 1 and Entry 8 will perform linear regression. We will be using 8.





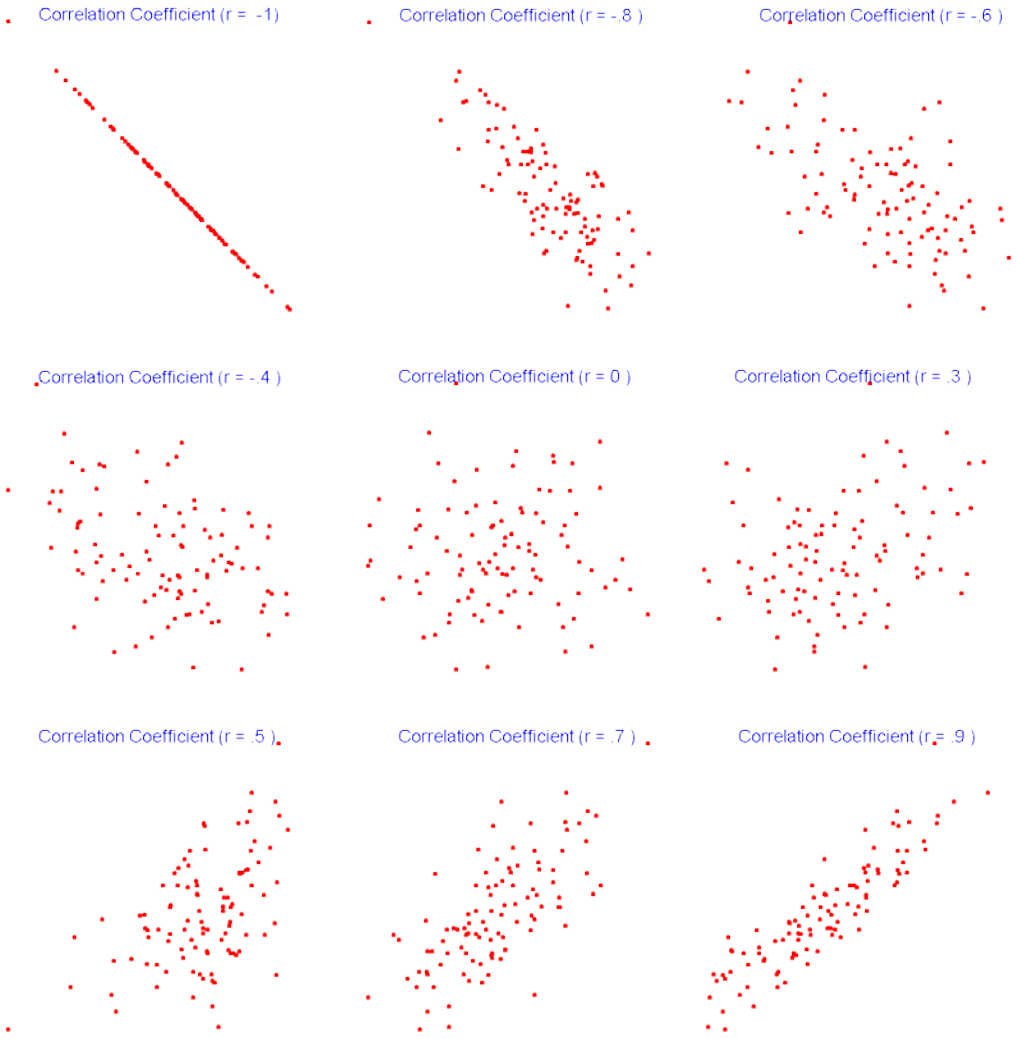
Put in the two lists and hit enter. r will be displayed on the next screen. Note that $r = 0.9668$, the same value found on the previous page.



- **Properties of r:**
  - The value of r does not depend on which of the two variables is labeled x (the explanatory variable).
  - The value of r does not depend on the unit of measurement for each value.
  - The value of r is between -1 and +1
  - The correlation coefficient is
    1. -1 only when all the points lie on a downward-sloping line (perfect negative relationship).
    2. +1 only when all the points lie on an upward-sloping line (perfect positive relationship).
  - The value of r is a measure of the extent to which x and y are linearly related.
  - Values of r near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 and toward -1 and +1.
  - Correlation measures the strength of only a LINEAR relationship between the two variables.
  - Correlation is not resistant – it is greatly affected by outliers.
  - Rough guidelines – value of r between -0.5 and $+0.5$ are considered weak, moderate values are from 0.5 to 0.8 (positive and negative) and values of r 0.8 and above are considered strong.

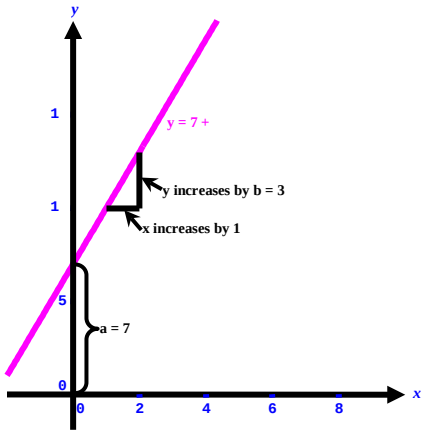
- Remember – even if $r = 0$, that does not mean there is *no* relationship between the variables, it just means that there is no LINEAR relationship. There could be another relationship relating the variables, like an exponential or parabolic relationship.
- **Ex:** For our value of r on the last example, 0.9668, we would say that there is a strong, positive linear association between the two variables. Note that in our sentence, we describe the strength, form and direction!!
- **Ex:** Hurricanes develop low pressure at their centers. This pulls in moist air, pumps up their rotation, and generates high winds. After looking at the data of the Maximum Wind Speed versus Central Pressure for 163 hurricanes that have hit the U.S. since 1851, the correlation coefficient was found to be $r = -0.879$. This means that there is a strong, negative linear association between the wind speeds of hurricanes and their central pressures.

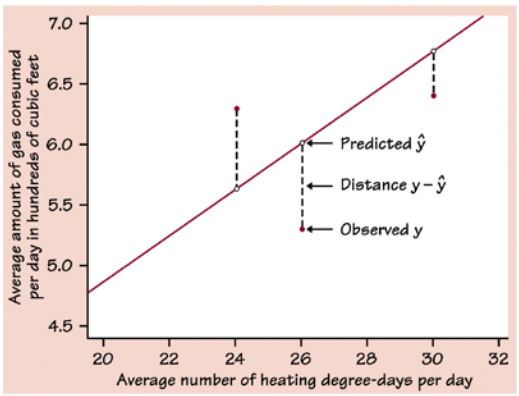
- **More Examples:**



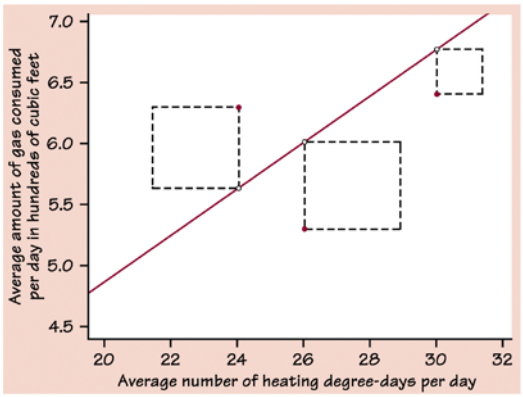
Syllabus Objectives: 1.15 – The student will determine the equation of the least squares line. 1.16 – The student will make predictions using the least squares regression line.

The Least Squares Regression Line

- We can model the relationship with a line and give its equation.
- The **linear model** is just an equation of a straight line through the data.
- A **regression line** is a line that describes how a response variable *y* changes as an explanatory variable *x* changes.
- **Regression Lines as Mathematical Models**
  - Form: $y = a + bx$
  - *b* is the **slope**, the amount by which *y* changes as *x* increases by one unit.
  - *a* is the **y-intercept**, the value of *y* when $x = 0$.
  - 

- We can use a regression line to predict the response *y* for a specific value of the explanatory variable *x*.
  - The estimate made from the model is called the **predicted value** and is written as $\hat{y}$ (read as y-hat). $\hat{y}$ is the prediction of *y* resulting from the substitution of a particular value into the equation.
  - The difference between the observed value and its associated predicted value is called the **residual**. The residual value tells us how far off the model's prediction is at that point. The residual is seen below as the vertical distance between the two values. $residual = observed\ value - predicted\ value$
  - 

- The **least-squares regression line** of *y* on *x* is the line that makes the sum of the squared residuals the smallest.
    - The line of best fit takes all of the sum of the squared residuals and makes them as small as possible. Here is a geometric representation of the squared (vertical distances) residuals.

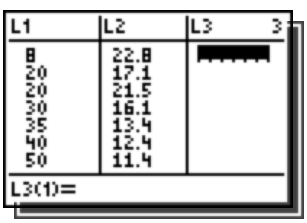


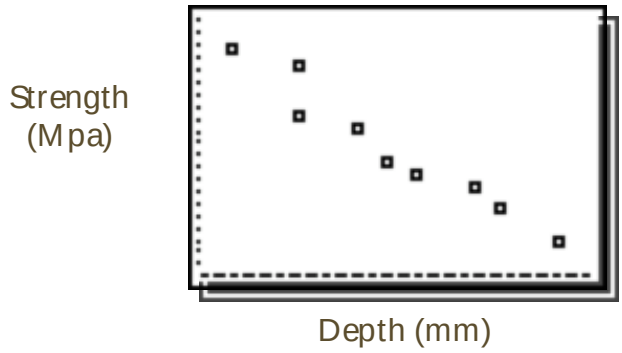
- **Equation** of the Least-Squares Regression Line (LSRL): $\hat{y} = a + bx$
    - The slope formula: $b = r\frac{s_y}{s_x}$, where *r* is the correlation coefficient and s_y and s_x are the standard deviation of the two variables.
    - The y-intercept formula: $a = \overline{y} - b\overline{x}$, where $\overline{x}$ and $\overline{y}$ are the two sample means. This formula works because the line always passes through the point $(\overline{x}, \overline{y})$.
    - Facts about the LSRL:
        - The distinction between explanatory and response variables is essential! If you mix them up, your LSRL will be wrong, but your correlation coefficient will be unchanged.
        - There is a close connection between correlation (*r*) and slope (*b*). Notice the formula - they will always have the same sign (positive or negative).
        - The LSRL can be found with the TI-84. Notice on the example on page 4, Stat Calc 8 will produce the LSRL.
        - We can use the regression line to predict the response *y* for a specific value of the explanatory variable *x*.
            - **Extrapolation** is the use of a regression line for prediction outside the range of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.

    - **Ex:** Consider the following data from the article, "The Carbonation of Concrete Structures in the Tropical Environment of Singapore". The explanatory variable is carbonation depth in concrete (mm) and the response variable is the strength of the concrete (Mpa)

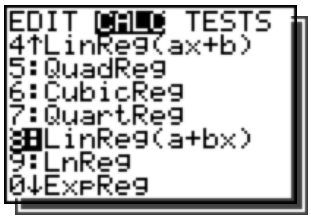
x	8	20	20	30	35	40	50	55	65
y	22.8	17.1	21.5	16.1	13.4	12.4	11.4	9.7	6.8

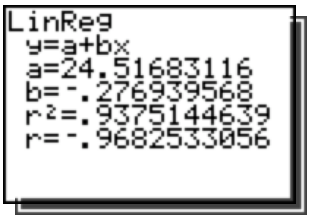
First, input the data into the calculator and draw the scatterplot.





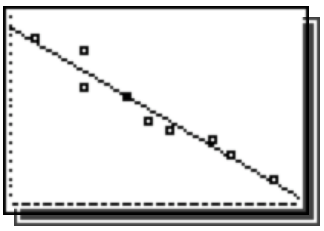
Interpret: There is a strong, negative linear relationship between depth of corrosion and concrete strength. As the depth increases, the strength decreases at a constant rate.
Next, find the equation of the LSRL that models the relationship between corrosion and strength.





$\hat{y} = 24.517 - 0.277x$, where x is the depth and $\hat{y}$ is the predicted strength.

$r = -0.968$, telling us that there is a strong, negative linear relationship between depth of corrosion and strength of concrete.
What does the slope tell us? For every increase of 1 mm in depth of corrosion, we predict on average, a 0.277 Mpa decrease in strength of the concrete.
Here the LSRL is drawn on the scatterplot.



Use the LSRL to find the predicted strength of concrete with a corrosion depth of 25 mm.

$\hat{y} = 24.517 - 0.277(25) = 17.592$

$\hat{y} = 17.592$ Mpa

What is the predicted strength of concrete with a corrosion depth of 40 mm?

$\hat{y} = 24.517 - 0.277(40) = 13.437$

$\hat{y} = 13.437$ Mpa

How does this prediction compare with the observed strength at a corrosion depth of 40 mm? The observed strength was 12.4 The prediction did not match the observation. That is, there was an "error" or "residual" between our prediction and actual observation. Our model predicted too high – the actual point is below the LSRL.

$$residual = observed - predicted$$

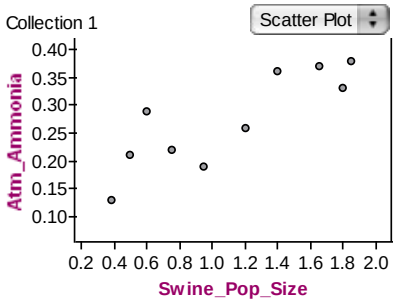
$$residual = 12.4 - 13.437 = -1.037$$

- **Coefficient of Determination (r^2)**
  - It is the proportion of variation in *y* that can be explained by the least-squares regression of *y* on *x*. It is the correlation coefficient squared.
  - Formula: $r^2 = \dfrac{SST - SSE}{SST}$ where $SST = \sum (y - \bar{y})^2$ and $SSE = \sum (y - \hat{y})^2$
    SST (SSTo in computer output) is the total sum of squares and *SSE* (SSResid in output) is the residual sum of squares. These can be found on computer outputs and you are not required to find them.
  - The closer the r^2 is to 1, the better the fit. If $r^2 = 1$, then 100% of the variation in *y* can be explained by the linear regression between the two variables. That means there would be a perfect fit, with no scatter around the line. All of the variance is accounted for by the model and none is left in the residuals at all. Of course, this is too good to be true. The coefficient determination will range from 0 to 1 and it measures the success of the regression model.
  - We can find the coefficient of determination by squaring *r* or it can be found on computer printouts and your TI-84.
  - **Ex:** From our previous example, notice the calculator screen shows: $r^2 = 0.9375$. This states that 93.75% of the variability in predicted strength can be explained by the LSRL on depth.

- **Linear Regression Example:** Animal-waste lagoons and spray fields near aquatic environments may significantly degrade water quality and endanger health. The National Atmospheric Deposition Program has monitored the atmospheric ammonia at swine farms since 1978. The data on the swine population size (in thousands) and atmospheric ammonia (in parts per million) for one decade are given below.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Swine Population	0.38	0.50	0.60	0.75	0.95	1.20	1.40	1.65	1.80	1.85
Atmospheric Ammonia	0.13	0.21	0.29	0.22	0.19	0.26	0.36	0.37	0.33	0.38

(a) Construct a scatterplot for these data.

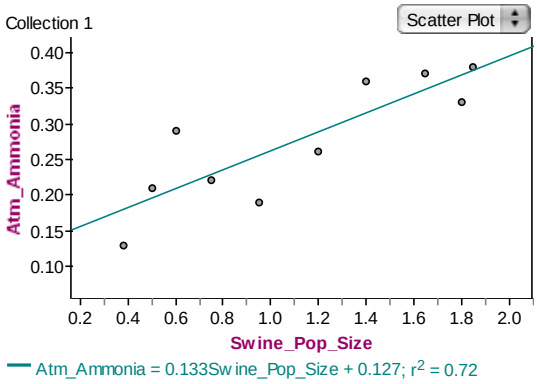


(b) The value for the correlation coefficient for these data is 0.85. Interpret this value.

There is a strong, positive, linear relationship between swine population size and atmospheric ammonia.

(c) Based on the scatterplot in part (a) and the value of the correlation coefficient in part (b), does it appear that the amount of atmospheric ammonia is linearly related to the swine population size? Explain.

Both the value of the correlation coefficient and the pattern in the scatterplot indicate that there is a positive linear relationship between the size of the swine population and atmospheric ammonia.



(d) What percent of the variability in atmospheric ammonia can be explained by swine population size?

$r^2 = 0.72$.

- **Computer Printouts**
  - **Minitab** output for regression. Often, the regression equation is given at the top. The slope *b* can also be found under the Coef column across from the explanatory variable (Distance). The *y*-intercept *a* is also under the Coef column across from "Constant". The coefficient of determination (r^2) is also given.

a

b

Regression Analysis: Standard Fare versus Distance

Least squares regression line

The regression equation is
Standard Fare = 10.1 + 0.150 Distance

Predictor	Coef	SE Coef	T	P
Constant	10.138	4.327	2.34	0.039
Distance	0.15018	0.01196	12.56	0.000

S = 6.803 R-Sq = 93.5% R-Sq(adj) = 92.9%

r^2

Analysis of Variance

SSResid

Source	DF	SS	MS	F	P
Regression	1	7298.1	7298.1	157.68	0.000
Residual Error	11	509.1	46.3		
Total	12	7807.2			

SSTo

- **Ex:** A scatterplot shows that there is a strong linear relationship between the average outside temperature (measured by heating degree-days) in a month and the average amount of natural gas that a household uses per day during the month. As the average number of heating degree-days per day increases, the average amount of gas consumed per day in hundreds of cubic feet also increases. The Minitab printout shows the following statistics.

The regression equation is:
$\hat{y} = 1.0892 + 0.1890x$.
The slope in this example says, that on average, each additional degree-day predicts consumption of 0.1890 more hundreds of cubic feet of natural gas per day.

The regression equation is
Gas Used = 1.09 + 0.189 D-days

Predictor	Coef	Stdev	t-ratio	p
Constant	1.0892	0.1389	7.84	0.000
D-days	0.188999	0.004934	38.31	0.000

s = 0.3389 R-sq = 99.1% R-sq(adj) = 99.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	168.58	168.58	1467.55	0.000
Error	14	1.61	0.11		
Total	15	170.19			

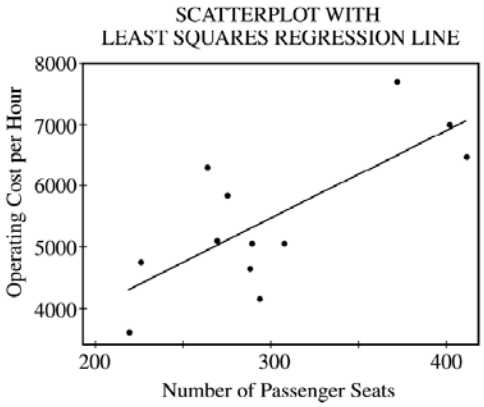
(a)

$r^2 = 0.991$
Over 99% of the variation in gas consumption is accounted for by the linear relationship with degree-days.

$r = \sqrt{0.991} = 0.995$. This shows us that there is a strong, positive linear relationship between gas used and D-days.

Using the equation to predict gas consumption at 20 degree-days, $\hat{y} = 1.0892 + (0.1890)(20) = 4.869$. For 20 degree-days, gas consumption would be about 4.869 hundreds of cubic feet.

- **Linear Regression Example:** Commercial airlines need to know the operating cost per hour of flight for each plane in their fleet. In a study of the relationship between operating cost per hour and number of passenger seats, investigators computed the regression of operating cost per hour on the number of passenger seats. The 12 sample aircraft used in the study included planes with as few as 216 passenger seats and planes with as many as 410 passenger seats. Operating cost per hour ranged between $3,600 and $7,800. Some computer output from a regression analysis of these data is shown below.



Predictor	Coef	StDev	T	P
Constant	1136	1226	0.93	0.376
Seats	14.673	4.027	3.64	0.005
S = 845.3	R-Sq = 57.0%		R-Sq (adj) = 52.7%	

(a) What is the equation of the least squares regression line that describes the relationship between operating cost per hour and number of passenger seats in the plane? Define any variables used in this equation.

$\hat{y} = 1136 + 14.673x$, where x = number of passenger seats and $\hat{y}$ = the predicted operating cost per hour.

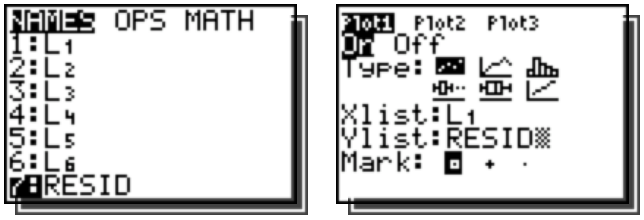
(b) What is the value of the correlation coefficient for operating cost per hour and number of passenger seats in the plane? Interpret this correlation.

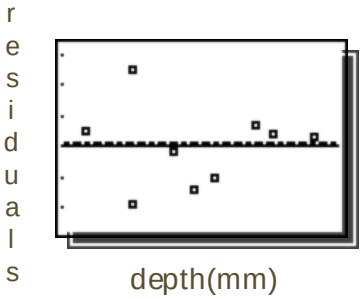
The value of the correlation coefficient is $r = \sqrt{0.570} = 0.755$. r is positive because the scatterplot shows a positive association and the slope is positive. There is a moderate, positive linear relationship between operating costs per hour and number of passenger seats.

(c) Suppose that you want to describe the relationship between operating cost per hour and number of passenger seats in the plane for planes only in the range of 250 to 350 seats. Does this line shown in the scatterplot still provide the best description of the relationship for data in this range? Why or why not?

No. The equation of the least-squares regression line is influenced by the three points in the upper right-hand corner and the two points in the lower left-hand corner of the scatterplot. The seven remaining points (with number of seats in the 250 to 350 range) would have a negative correlation. Hence, the slope of the recalculated least-squares regression line is negative.

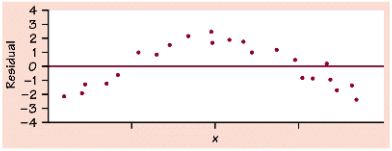
Syllabus Objectives: 1.17 – The student will analyze residual plots for patterns. 1.18 – The student will analyze the effect of outliers and influential points on the least squares regression line.

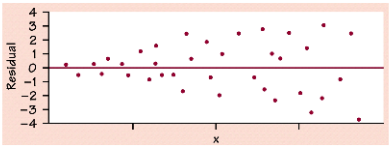
- A **residual plot** is a scatterplot of the regression residuals against the explanatory variable, *x*.
  - Residual plots help us to assess the fit of a regression line.
  - We are looking for NO PATTERN or CURVATURE. Uniform or random scatter in the residual plot tells us that a linear model is appropriate.
  - If there is curvature, increasing or decreasing spread, or lots of points with large residuals, this is an indicator that the linear regression is not a good fit for the data.
  - The residuals are found for each data point and plotted on the vertical axis. The explanatory variable is plotted on the horizontal axis. No need to do by hand. Each time Linear Regression (Stat, Calc, 8) is performed on the graphing calculator, the list named Resid is created. This list contains all of the residuals. Find this list under the named lists and plot the scatterplot.
  - **Ex:** Returning to our running example of corrosion depth and strength of concrete. Find the named list Resid and draw the scatterplot.
  - 



    There appears to be no pattern in the residual plot. The LSRL may be our best prediction model.

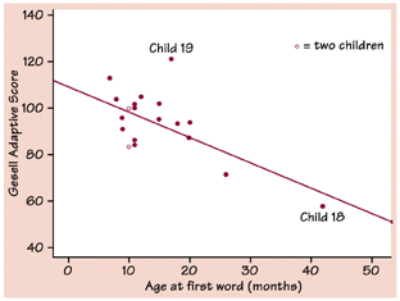
  - Examples of residual plots where a straight line may not be the best model. The first shows curvature, the second shows a fanned pattern.

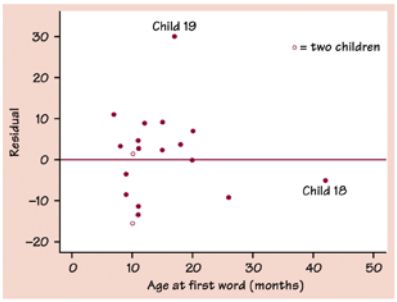




- **Outliers and Influential Observations**
  - An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.
  - An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the *x* direction of a scatterplot are often influential for the LSRL. We say that a point is influential if omitting it from the analysis gives a very different model.

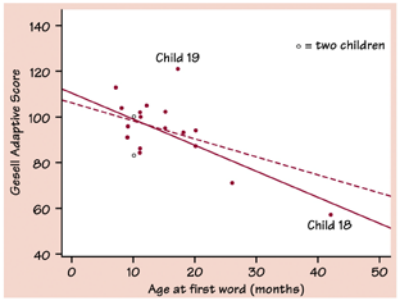
- **Ex:** Does the age at which a child begins to talk predict later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive score. **Note: Child 18 and Child 19.**





We can see from the residual plot that Child 19 is an outlier with a very large residual. Child 18 is an influential observation that does not have a large residual. If Child 18 is removed from the data, it greatly affects the LSRL.

This figure shows two least-squares regression lines. The solid line is calculated from all the data. The dashed line is calculated leaving out Child 18. Child 18 is influential because it moves the regression line quite a bit.

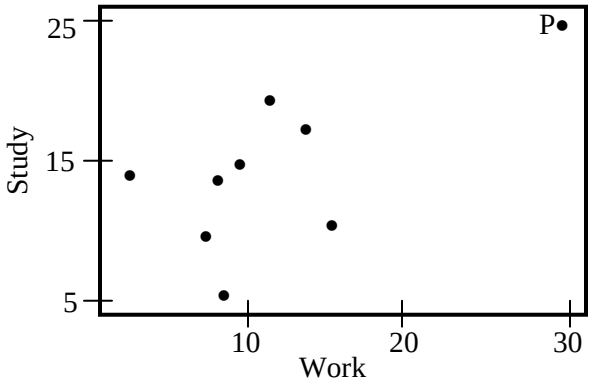


- **Influential Observation Example:** A simple random sample of 9 students was selected from a large university. Each of these students reported the number of hours he or she had allocated to studying and the number of hours allocated to work each week. A least squares regression was performed and part of the resulting computer output is shown below.

Predictor	Coef	StDev	T	P
Constant	8.107	2.731	2.97	0.021
Work	0.4919	0.1950	2.52	0.040

S = 4.349 R-Sq = 47.6% R-Sq (adj) = 40.1%

The scatterplot to the right displays the data that were collected from the 9 students.



(a) After point P, labeled on the previous graph, was removed from the data, a second linear regression was performed and the computer output is shown below.

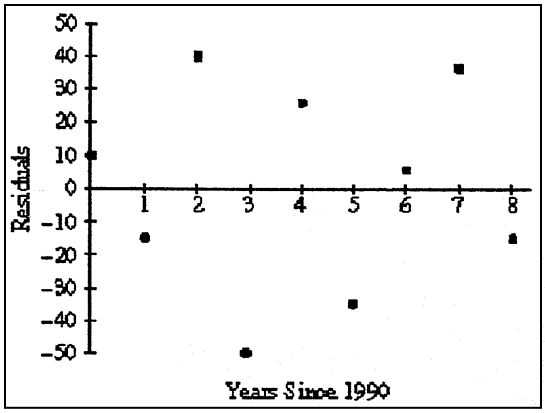
Predictor	Coef	StDev	T	P
Constant	11.123	3.986	2.79	0.032
Work	0.1500	0.3834	0.39	0.709

S = 4.327 R-Sq = 2.5% R-Sq (adj) = 0.0%

Does point P exercise a large influence on the regression line? Explain.

The point P does have a large influence on the regression line. When P is removed from the data set, the slope of the line changes from 0.4919 to 0.1500, the intercept changes from 8.107 to 11.123, and the value of R^2 drops from 47.6% to 2.5%. Also, the shape is significantly different from 0 when the point P is included in the data set and is not significantly different from 0 when the point P is excluded from the data set.

- **Ex:** Lydia and Bob were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990−1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least-squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s = 33.43				

(a) Is a line an appropriate model to use for these data? What information tells you this?

Yes. The residual plot shows no pattern, indicating a linear model is appropriate.

(b) What is the value of the slope of the least squares regression line?
Interpret the slope in the context of this situation.

Slope = 233.517 aircraft/year
On average, the number of commercial aircraft flying in the U.S. increased by approximately 233.517 each year.

(c) What is the value of the intercept of the least squares regression line?
Interpret the intercept in the context of this situation.

Intercept = 2939.93 aircraft.
Predicted number of commercial aircraft that were flying in 1990 (since x=0 corresponds to year 1990) was 2939.93.

(d) What is the predicted number of commercial aircraft flying in 1992?

For 1992, x = 2, so predicted number of commercial aircraft flying is $2939.93 + 233.517(2) = 3406.964$ *aircraft.*

(e) What is the actual number of commercial aircraft flying in 1992?

From the residual plot, the residual for 1992 is +40, $40 = actual - predicted$.
$actual = 3406.964 + 40 = 3446.964$ *aircraft. Since actual number flying must be an integer, actual must have been 3447.*

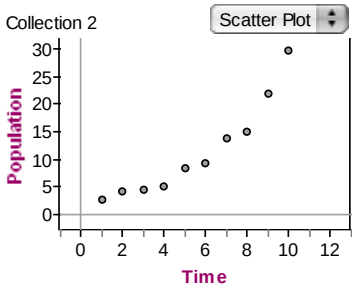
Syllabus Objective: 1.19 – The student will transform bivariate data to achieve linearity, including logarithmic and power transformations.

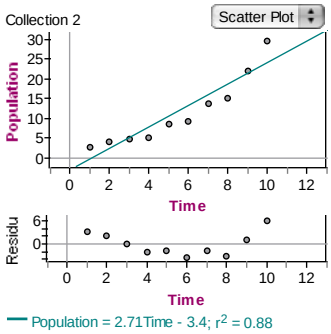
- **Transforming to achieve linearity**
  - Applying a function such as a logarithm or power to a quantitative variable is called **transforming** or **reexpressing** the data.
  - This helps us to straighten nonlinear patterns. Once the curved data is straightened, we can use the tools of linear regression to summarize and analyze our data.
  - To decide which transformation to use, the process is often one of trial and error. There are two major types – after the model is recognized, the transformation process becomes much simpler.
  - Steps: Make a scatterplot of the data and also find the residual plot. If the scatterplot is curved or if the residual plot is curved, a linear model is not appropriate and the data must be transformed. Make one of the transformations described below. After the data has been reexpressed, graph the scatterplot to check for linearity and also check the residual plot for random scatter. If both illustrate the new model is linear, run the linear regression on the new data. When predicting, apply the inverse log operation to isolate the variables.

- **Exponential growth**
  - In linear growth, a fixed increment is *added* to the variable in each equal time period. **Exponential growth** or **decay** occurs when a variable is *multiplied* by a fixed number in each equal time period.
  - To straighten an exponential model, find the logarithm of the *y*-values. Ln or Log may be applied.

  - **Ex. of Exponential growth:** A researcher observes the growth of a particular bacteria and records the following results:

Time (hr)	1	2	3	4	5	6	7	8	9	10
Population (in thousands)	2.61	4.19	4.65	5.27	8.46	9.35	13.74	15.06	21.98	29.72

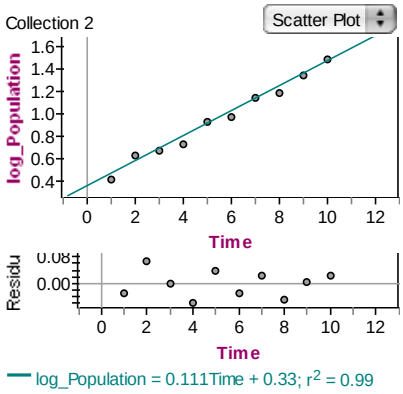




The scatterplot **does not** follow a linear pattern. Although r^2 is high, the curvature in the scatterplot and the residual plot tell us that a linear model is not appropriate. The data appear to follow an exponential model so a transformation of $y \to \log y$ is appropriate.

Time (hr)	1	2	3	4	5	6	7	8	9	10
Population	2.61	4.19	4.65	5.27	8.46	9.35	13.74	15.06	21.98	29.72
Log Pop	0.4166	0.6222	0.6674	0.7218	0.9274	0.9708	1.138	1.1778	1.342	1.473

After taking the log of the *y*-values, note that the scatterplot does follow a linear pattern and there is no pattern apparent in the residual plot. The computer output shows the least-squares regression line and a very high coefficient of determination. Now the line can be used for prediction. Suppose we want to predict the population at 15 hours.



$$\log \hat{y} = 0.33 + 0.111x$$

$$\log \hat{y} = 0.33 + 0.111(15) = 1.995$$

$$\log \hat{y} = 1.995 \Rightarrow 10^{1.995} = 98.855$$

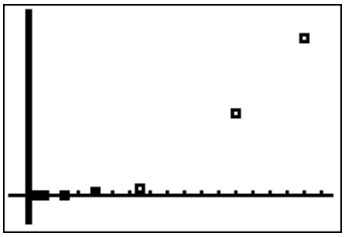
$$\hat{y} \approx 98.855$$

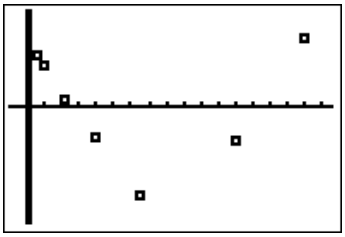
Our prediction for the population of this bacteria at 15 hours is 98.855 thousands of bacteria, or approximately 98,855 bacteria.

- **Power model**
  - Examples of the power (square, cube) model – weight, volume, area
  - To produce a linear relationship from a power law model, apply the logarithm transformation to *both* variables. Again, natural logarithms (ln) or Base 10 logs (log) may be applied.

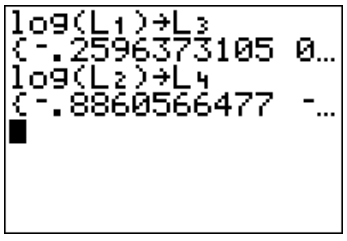
  - **Ex. of Power Model:** Scientists looked at the heart weight (in grams) of 7 mammals and the length of the cavity of the left ventricle of their hearts (in centimeters) to discover if there was some kind of relationship.

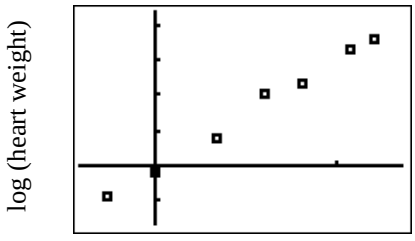
Mammal	Mouse	Rat	Rabbit	Dog	Sheep	Ox	Horse
Cavity Length (cm)	0.55	1.0	2.2	4.0	6.5	12.0	16.0
Heart Wt. (gms)	0.13	0.64	5.8	102	210	2030	3900

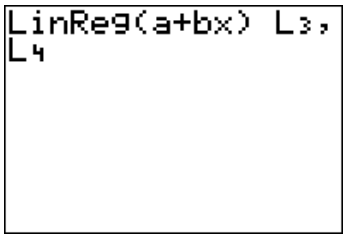


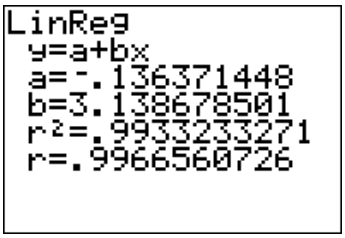


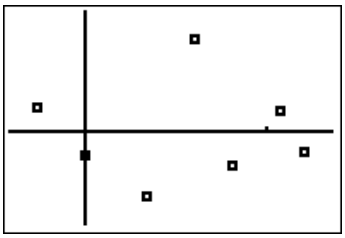
The scatterplot (above left) shows curvature and the residual plot (above right) also shows curvature. A linear model is clearly not appropriate. Because weight implies this might be a power model, apply the logarithm to both variables. The log of cavity length was stored in L_3 and the log of heart weight was stored in L_4. A scatterplot of the new transformed variables shows a very linear pattern.











Residual plot

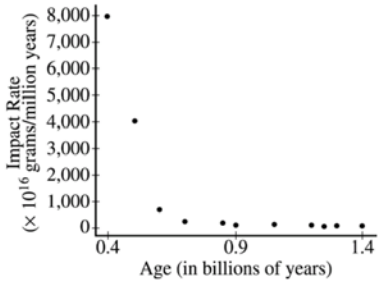
The new residual plot shows random scatter implying the power model is the appropriate model. The least-squares regression line for the new model is $\log \hat{y} = -0.1364 + 3.139 \log x$. We can now use the model to make predictions. To predict the heart weight for a mammal with a cavity length of 10 centimeters:

$\log \hat{y} = -0.1364 + 3.139 \log(10) = -0.1364 + 3.139(1) = 3.0026$

$\log \hat{y} = 3.0026 \Rightarrow 10^{3.0026} = \hat{y} \approx 1006.00$

Our prediction for the heart weight of a mammal with a left ventricle cavity length of 10 cm is approximately 1006 grams.

- **Example:** The Earth's Moon has many impact craters that were created when the inner solar system was subjected to heavy bombardment of small celestial bodies. Scientists studied 11 impact craters on the Moon to determine whether there was any relationship between the age of the craters (based on radioactive dating of lunar rocks) and the impact rate (as deduced from the density of the craters). The data are displayed in the scatterplot.



(a) Describe the nature of the relationship between impact rate and age.

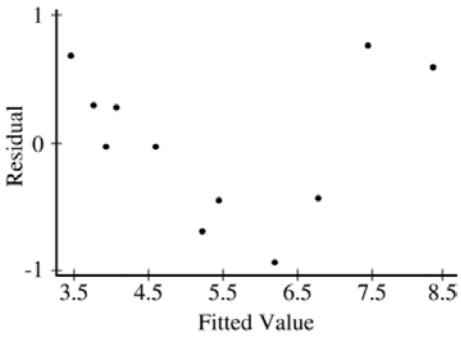
There is a strong nonlinear relationship between impact rate and age. Impact rate declines rapidly with age over the age range from 0.4 to about 0.7 billion years, and then seems to level out.

Prior to fitting a linear regression model, the researchers transformed both impact rate and age by using logarithms. The following computer output and residual plot were produced.

Regression Equation: ln(rate) = 4.82 − 3.92 ln(age)

Predictor	Coef	SE Coef	T	P
Constant	4.8247	0.1931	24.98	0.000
ln(age)	−3.9232	0.4514	−8.69	0.000

S = 0.5977 R-Sq = 89.4% R-Sq (adj) = 88.2%



(b) Interpret the value of r^2.

89.4% of the variability in ln(impact rate) can be explained by a linear or straight line, relationship between ln(impact rate) and ln(age).

(c) Comment on the appropriateness of this linear regression for modeling the relationship between the transformed variables.

There is a noticeable curved pattern in the residual plot, which indicates that the linear model is not the best choice for describing the relationship between ln(impact rate) and ln(age).