Gamma Iota Sigma will be bringing people from the Society of Actuaries and they will be talking about their careers as actuaries. They will also be discussing what the Society of Actuaries can provide for students who are interested in Actuarial Sciences. The location, time and date are:

Location: CEMO 100D
Time: 6:30 - 7:30
Date: Wednesday 11/16

Math 3339

Section 27204
MWF 10-11:00am AAAud 2

Bekki George
bekki@math.uh.edu
639 PGH

Office Hours:
M & Th noon – 1:00 pm & T 1:00 – 2:00 pm
and by appointment

The Simple Linear Regression Model

A response variable (dependent) measures the outcome of a study.

An explanatory variable (independent) attempts to explain the observed outcomes.

The most common graphical display used to study the association between two variables is called a scatter plot.

The bivariate data given below relate the high temperature reached on a given day and the number of cans of soft drinks sold from a particular vending machine.

Temp.	Cans	Temp.	Cans	Temp.	Cans
70	30	98	59	90	53
75	31	72	33	95	56
80	40	75	38	98	62
90	52	75	32	91	51
93	57	80	45	98	58

To graphically analyze the data, we can display the data on a 2-dimensional graph. Which is the dependent (x) variable and which is the independent (y) variable?

Using R:
>plot(temp,cans)

To interpret a scatter plot look for the *direction*, *form* and *strength.*

Direction: Positive association or negative association

Two variables are said to be **positively** related if larger values of one variable tend to be associated with larger values of the other.

Two variables are said to be **negatively** related if larger values of one variable tend to be associated with smaller values of the other.

Form : Shape

The form of a scatter plot refers to its shape. Most of what we will see are linear forms.

Strength: How closely the points follow the shape

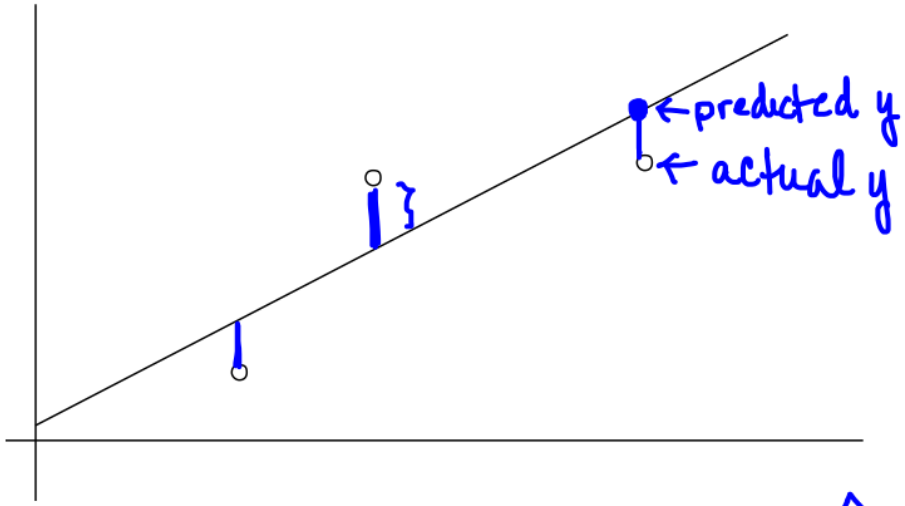
The closer the points fall in to a straight line, the stronger we say the relationship is.

You may introduce categorical variables into a scatter plot by using a different symbol for the dots of various categorical variables. For example, solid dots for males, open dots for females.

Least-Squares Regression

The least-squares regression line is a mathematical model for the data that allows us to predict the value of y for a given value of x.

We want a regression line that makes the vertical distances of the points in a scatter plot from the line as small as possible.



error = residual = act. y − pred y

The least squares regression line: $\hat{y} = a + bx$

b: Slope

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

$$b = r\frac{s_y}{s_x}$$ r = correlation

$$a = \bar{y} - b\bar{x}$$ mean of all y's; mean of all x's

note the pt. $(\bar{x}, \bar{y})$ is always on LSRL

Using R and the data above:

>temp=c(…)

>cans=c(….)

>lin_mod=lm(cans~temp) this command will always have y~x

>lin_mod

$y \sim x$

$\hat{y}_{cans} = -46.171 + 1.086x_{temp}$

we will be given our y-intersept and slope

>plot(temp,cans) this will give the scatter plot

>abline(lin_mod) this will show the regression line on the scatter plot

Let's investigate some properties of the least-squares regression line:

Slope: the slope is the rate of change, the amount of change in $\hat{y}$ when x increases by 1 unit. In this example, a slope of 1.086 says that each additional degree-day predicts consumption of 1.086 more ~~hundreds of cubic feet of natural gas~~ cans of soda per day.

Correlation Coefficient, *r*

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The correlation r is an average of the products of the standardized values of x and y for n data pieces.

Facts about Correlation:

$r > .8 \Rightarrow$ strong pos. lin. assoc
$r < -.8 \Rightarrow$ " " neg " "

1. Positive r indicates positive association and negative r indicates negative association between variables.
2. r is always between –1 and 1 $\quad -1 \le r \le 1$
3. Correlation is strongly influenced by outliers.

.5 - .8 moderate
< .5 weak

The coefficient of determination is a measure that allows us to determine how certain one can be in making predictions with the line of best fit. It measures the proportion of the variability in the dependent variable that is explained by the regression model through the independent variable.

- The coefficient of determination is obtained by squaring the value of the correlation coefficient.
- The symbol used is r^2
- Note that $0 \le r^2 \le 1$
- r^2 values close to 1 would imply that the model is explaining most of the variation in the dependent variable and may be a very useful model.
- r^2values close to 0 would imply that the model is explaining little of the variation in the dependent variable and may not be a useful model.

R command for the correlation is cor(x,y)

Residuals

Residuals are just errors. A residual is the difference between an actual observed y value and the corresponding predicted y value.

Residual = error = (observed – predicted) = $(y - \hat{y})$ actual y - predicted y

Plots of residuals may display patterns that would give some idea about the appropriateness of the model. If the functional form of the regression model is incorrect, the residual plots constructed by using the model will often display a pattern. The pattern can then be used to propose a more appropriate model.

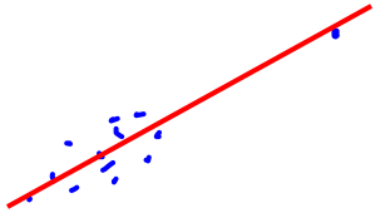
To view residual plot >plot(lin_mod)

Outliers vs. Influential Points

An outlier is a value that is well separated from the rest of the data set. An outlier will have a large absolute residual value.

An observation that causes the values of the slope and the intercept in the line of best fit to be considerably different from what they would be if the observation were removed from the data set is said to be influential.

Outlier



Influential

Example: Televisions and Life Expectancy

Country	Life Exp.	Per TV (people)	Country	Life Exp.	Per TV
Angola	44	200	Mexico	72	6.6
Australia	76.5	2	Morocco	64.5	21
Cambodia	49.5	177	Pakistan	56.5	73
Canada	76.5	1.7	Russia	69	3.2
China	70	8	S. Africa	64	11
Egypt	60.5	15	Sri Lanka	71.5	28
France	78	2.6	Uganda	51	191
Haiti	53.5	234	U.K.	76	3
Iraq	67	18	U.S.	75.5	1.3
Japan	79	1.8	Vietnam	65	29
Madagascar	52.5	92	Yemen	50	38

a) Which of the countries listed has the fewest people per television set? US
Which has the most? What are those numbers?
Haiti

b) Use the ~~calculator~~ to produce a scatter plot. Does there appear to be an association?
moderate neg. association ($r = -.6$)

c) Have the calculator compute the value of the correlation coefficient between life expectancy and people per television. r = -.6

d) Since the association is so ~~strongly~~ moderately negative, one might conclude that simply sending television sets to the countries with lower life expectancies would cause their inhabitants to live longer. Comment on that argument.

e) If two variables have a correlation close to +1 or –1, indicating a strong linear relationship, does it follow that there must be a cause-and-effect relationship between them?

This example illustrates a very important distinction between association and causation. Two variables may be strongly associated without a cause-and-effect relationship existing between them. Often the explanation is that both variables are related to a third variable not being measured; this variable is often called a *lurking* or *confounding* variable.

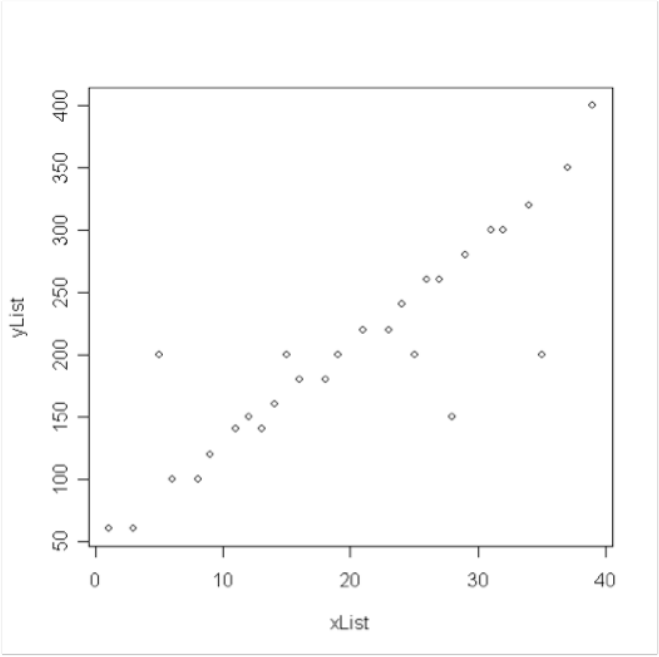
f) In this case, suggest a confounding variable that is associated with both a country's life expectancy and the prevalence of televisions in the country.

More on R commands:

Enter your lists:
> xList=c(..)
> yList=c(..)

To do a scatterplot, the command is plot(x, y)
>plot(xList, yList)

The correlation is found by using cor(x,y)
> cor(xList, yList)
[1] 0.8779736



To find the LSRL, we will use the lm() function. The parameter for this function is called the **model formula**.
The model formula is y ~ x (read as y is modeled by x).

```
> yp=lm(yList~xList)
> yp
```

linear model

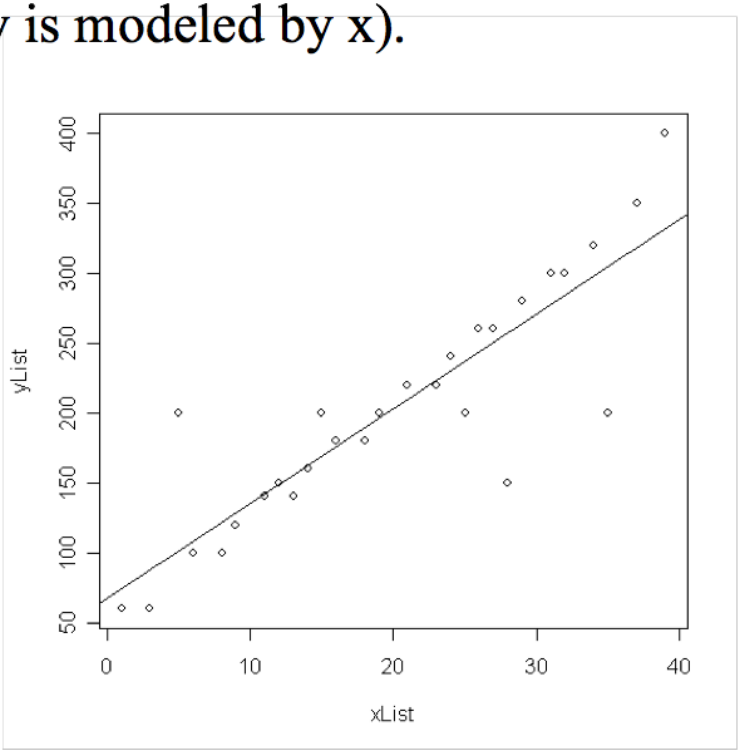
```
Call:
lm(formula = yList ~ xList)

Coefficients:
(Intercept)       xList
     67.283       6.784
```

$\hat{y} = a + bx$

a b

So, the LSRL is $\hat{y} = 67.283 + 6.784x$



The function abline() after the plot() command will make the LSRL appear on the graph:

```
> plot(xList, yList)
> abline(yp)
```

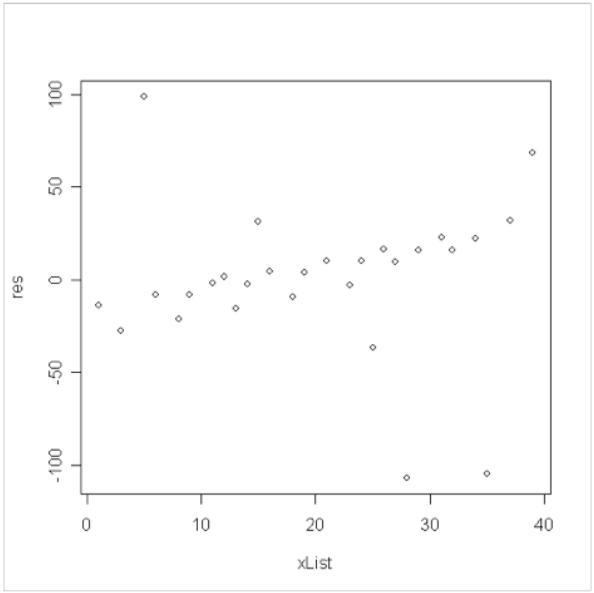
residuals() will find and list all residual values

```
> res=residuals(yp)
> res
        1         2         3         4         5
-14.067204 -27.636128  98.794948  -7.989515 -21.558439
```

Want no pattern in residuals

……..

```
> plot(xList, res)
```



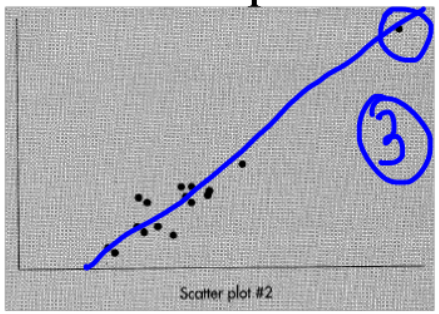
You can see many plots by using plot with your model.

```
> plot(yp)
```

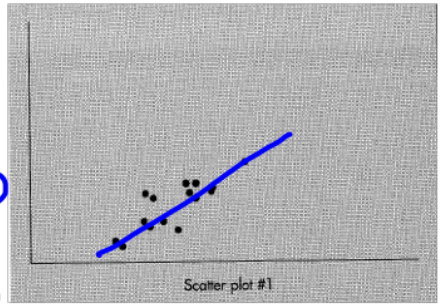
lin model

Popper 30

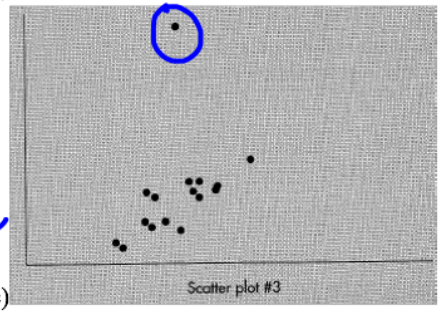
Match each plot to the equation of the LSRL and it's r



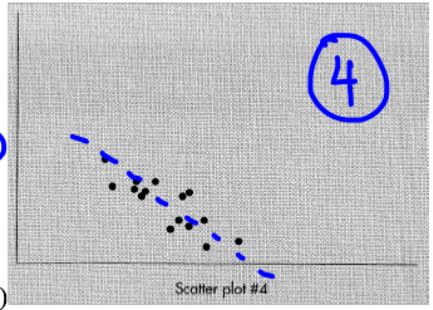
a)



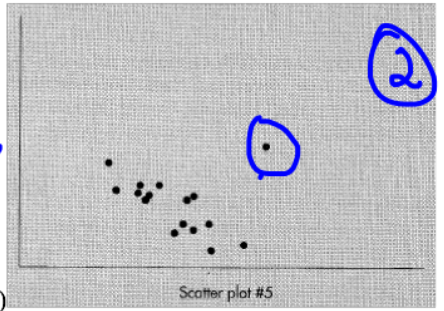
b)



c)



d)



e)

1. $r = 0.83,\ y = 1.4x - 2.1$

2. $r = -0.31,\ y = -0.5x + 7.8$

3. $r = 0.96,\ y = 1.4x - 2.1$

4. $r = -0.83,\ y = -1.4x + 11.8$

5. $r = 0.41,\ y = 1.4x - 1.4$