

3.3 Least Squares Regression

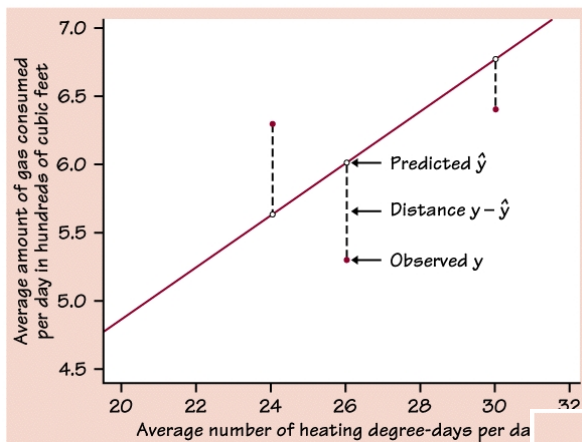
Quantitative Bivariate data:

- Create a scatterplot
 - Interpret (form/direction/strength - is it **LINEAR**?)
 - number summary (\bar{x} , \bar{y} , s_x , s_y , r)
 - create a mathematical model

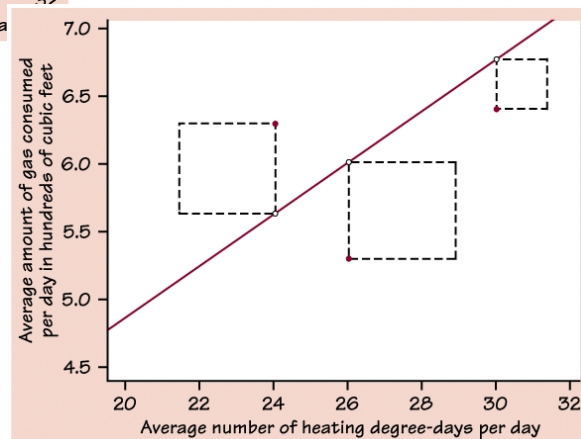
Least Squares Regression Line (LSRL) - a mathematical model for linear data.

- based on minimizing the vertical distance (error) between all of the predicted y values (\hat{y} or y-hat) and the actual y values (y)
- error = observed - predicted ($y - \hat{y}$)
- use a slope/intercept equation to model linear trends
- LSRL always passes through the point (\bar{x} , \bar{y}).

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



Least Squares Demo
LSRS Applet (NCTM)
Sketchpad LSRS applet



Finding the LSRL equation:

need slope (b) and intercept (a) to make the equation:

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

$$\hat{y} = b_0 + b_1x$$

you will use the following formulas for a and b with the same numbers you generated when you performed 2-var stats on your calculator.

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

$$b_1 = r \frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Let's try an easy one:

Use the Archaeopteryx data from 3.24 (on P142)

Femur: 38 56 59 64 74
Humerus: 41 63 70 72 84

I'll give you $r = .994$ for this one

```
2-Var Stats
x=58.2
Σx=291
Σx²=17633
Sx=13.19848476
σx=11.80508365
↓n=5
```

```
2-Var Stats
↑y=66
Σy=330
Σy²=22790
Sy=15.89024858
σy=14.2126704
↓Σxy=20040
```

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

$$a = \bar{y} - b\bar{x}$$

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

Now let's use your calculator to generate the equation:

Before you start - go to **2nd | 0** (Catalog) and select "Diagnostics On"

Now - with L_1 as your explanatory (x) variable and L_2 as your response (y) variable, select **STAT | CALC | 8** and then enter.

(Calc defaults to L_1 and L_2)

```

EDIT [2nd] [0] TESTS
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
  
```

```

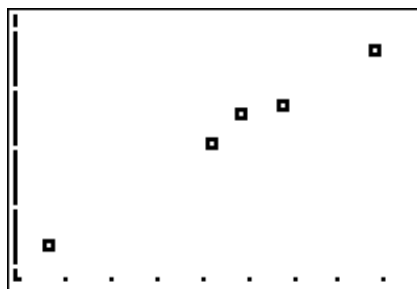
LinReg(a+bx)
  
```

```

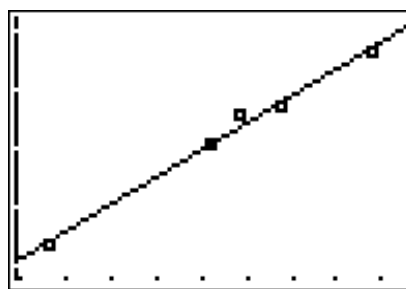
LinReg
y=a+bx
a=-3.659586682
b=1.196900115
r2=.9883313819
r=.9941485714
  
```

You can use these a and b values to write your regression equation.

You will need to be able to create a regression equation from raw data (using the calculator) and from a numerical summary (using the formulas).



Scatterplot of our data



and with $\hat{y} = -3.6596 + 1.1969X$

When you are plotting a scatterplot with regression line by hand, use the equation to find two predicted values near the ends of the x range, plot those points and then draw the line through them.

r^2 (the coefficient of determination): More detail on this later, (read pp 158-162 for a great example).

for now - the r^2 for this example was about .988 - so we would say that 98.8% of the variation in y values can be explained by regression (straight line dependence) of y on x. The stronger the correlation (r), the higher percent of variation can be explained by r^2 .

Residuals and Residual Plots

The "error" in prediction from a regression equation is called a residual and is designated as:

$$\text{residual} = y - \hat{y}$$

Residuals measure the vertical distance between every y value in a set of data and the predicted value \hat{y} that can be calculated with the LSRL equation.

It is the sum of the squared residuals that helps us get our LSRL equation.

residuals are:

- + when above the LSRL
- when below it
- add up to "0" (or very close)
- the mean of the residuals = 0

Residual Plots

- A scatterplot of residuals against explanatory variable ($x, y - \hat{y}$)
- Help us assess how well a regression line fits the data
- Can be created easily in your calculator:

L1	L2	L3	3
38	41	-----	
56	63		
59	70		
64	72		
74	84		
-----	-----		
L3=L2-Y1(L1)			

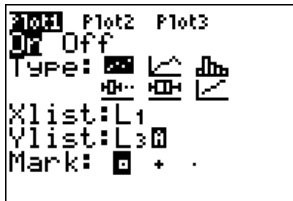
In the header for L3, type in the formula for $y - \hat{y}$ using the regression function stored in Y-VARS Y_1

L1	L2	L3	3
38	41	-.8226	
56	63	-.3668	
59	70	3.0425	
64	72	-.942	
74	84	-.911	
-----	-----	-----	
L3={-.822617680...			

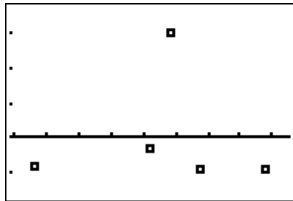
The residual values for each observed value y are now in List 3.

sum(L3)	1E-12
---------	-------

When you add all of the residuals together, the sum is virtually zero.

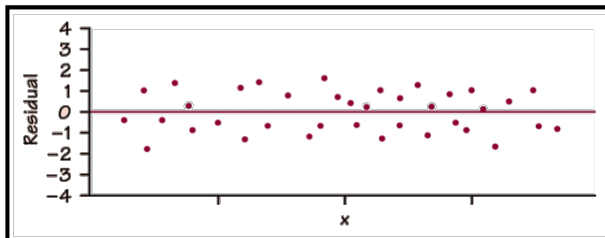


To plot the residuals, use the explanatory values in L₁ as your X list and the residual values in L₃ as your Y list and ZOOM 9 to see your plot.

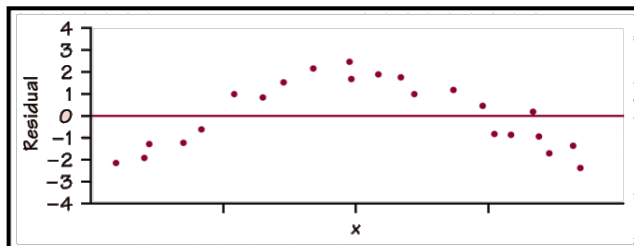


If the regression equation is a good model for your data, then the points should look scattered in an unstructured way around the "0" line, which is the mean of the residuals

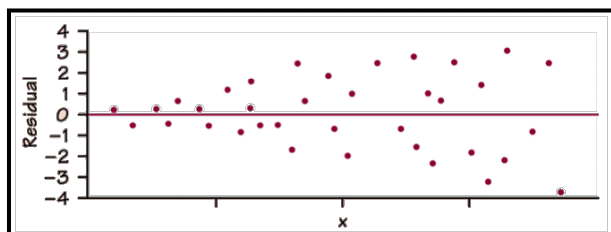
If a residual plot takes on any type of structure or shape, then it can be an indicator of poor linear fit.



Good Model - uniform scatter no pattern or structure



Curved Pattern - straight line is not a good model



increasing spread - predictions lose accuracy as x gets bigger

Influential Observations and outliers

Outliers:

- outside the overall pattern of other observations
- can be in any direction from the cluster
- do not necessarily affect the LSRL

Influential Observations in a scatterplot:

- An observation is influential if removing it would change the LSRL significantly.
- Vertical outliers tend to shift the LSRL up and down
- Horizontal outliers (in the x direction) tend to be influential to an LSRL

PoS Applet