Non-Linear Relationships

Section 5.4

Cathy Poliak, Ph.D.
cathy@math.uh.edu
Office in Fleming 11c

Department of Mathematics
University of Houston

Lecture 14 - 2311

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Popper Set Up

- Fill in all of the proper bubbles.
- Make sure your ID number is correct.
- Make sure the filled in circles are very dark.
- This is popper number 10.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Calcuating LSLR

- Given the statistics: $\bar{x}, \ \bar{y}, \ s_x, \ s_y,$ and r. Use the following to calulate $\hat{y} = a + bx$.

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

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

- Given the data set in R use:
  lm(y $\sim$ x)

- The residual is the difference between the actual y-value and the predicted y-value

$$resid = y - \hat{y}$$

- The coefficient of determination is the sqared value of the correlation coefficent; R^2. This is the percent (fraction) of the variation of y that can be explained by the LSLR.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

The following data was collected comparing score on a measure of test anxiety and exam score:

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

We will use R to:

- Construct a scatterplot.
- Find the LSRL and fit it to the scatterplot.
- Find r and r^2.
- Does there appear to be a linear relationship between the two variables? Based on wha you found, would you characterized the relationship as positive or negative? Strong or weak?
- Draw the residual plot.
- What does the residual plot reveal?
- Answer the question: Is it reasonable to conclude that test anxiety caused poor exam performance?

```
> anxiety=c(23,14,14,0,7,20,20,15,21)
> exam=c(43,59,48,77,50,52,46,51,51)
> plot(anxiety,exam)
> lm(exam~anxiety)
```

← scatter plot

← LSRL

```
Call:
lm(formula = exam ~ anxiety)

Coefficients:
(Intercept)     anxiety
     68.838      -1.064
```

← $\hat{y} = 68.838 - 1.064x$

$x = 15$ $\quad \hat{y} = 68.838 - 1.064(15) = 52.878$

Residual for $x = 15$: $y = 51$ $\quad$ residual $= 51 - 52.878$

$= -1.878$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

```
> cor(anxiety,exam)
[1] -0.7877352
> cor(anxiety,exam)^2
[1] 0.6205267
```

correlation coefficient

negative, somewhat strong relationship.

Coefficent of determination

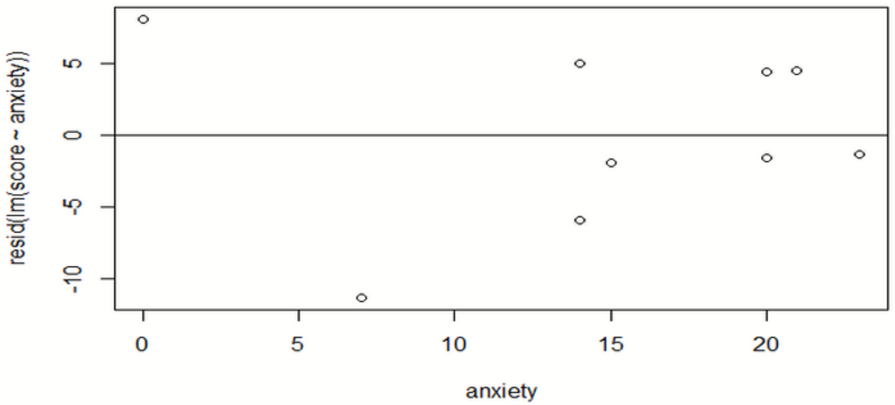
About 62% of the variation in the exam scores can be explained by this LSRL.

To add on the line to the scatterplot

```
> plot(anxiety,exam)
> abline(lm(exam~anxiety))
>
```

Residual Plot

```
> plot(anxiety,resid(lm(exam~anxiety)))
> abline(0,0)
```



Caution about correlation and regression

- An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of sacatterplot have large regression residuals.
- **Influential Observations**: a data point whose removal causes the regression equation (and the line) to change considerably. Points that are extreme in the x direction of a scatterplot are often influential for the least-squares regression line.
- **Data points do not appear to be scatter about a straight line**: First draw a scatter diagram if the points do not appear to be scattered about a straight line, do not determine a regression line.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Caution about correlation and regression

- **Extrapolation**: Using the regression equation to make predictions for values of the predictor variable outside the range of the observed values of the predictor variable. (Grossly incorrect predictions can result from extrapolation.)
- **Lurking Variable**: A variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
- **Association does not imply causation.**

Popper 10 Questions

$b = r\frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

$b = 0.93257\left(\frac{28.0842}{25.87901}\right)$ $a = 71.09433 - 1.012(71.17697)$

The following are the statistics to calculate a least squares line (LSLR) to predict the final grade (**score**) by the average **quiz** score. **Score**: mean = 71.09433, sd = 28.0842; **Quiz**: mean = 71.17697, sd = 25.87901; **Correlation**: r = 0.93257

1. Determine the LSLR to predict the final **score** (y) based on the **quiz** (x) score.

   a. $\hat{y} = -0.939326 + 1.012x$ b. $\hat{y} = 10.0826 + 0.8593x$
   c. $\hat{y} = 70.913$ d. $\hat{y} = 1 + 0.85x$

2. A person has a **quiz** average for the semester at 86, has a final score of 93. Determine the residual for this student. $x = 86$ $y = 93$

   $\hat{y} = 86.1$

   $\text{resid} = y - \hat{y} = 93 - 86.1$

   a. 6.9 b. 86.1 c. -6.9 d. 0

3. Determine the coefficient of determination.

   a. 0.93257 b. 0.8967 c. 0.9215 d. 1.0852

$r^2 = 0.93257^2$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

Can we predict the people per square mile based on year? The following is the population int he city of Houston from 1900 to 2010.

Year	Population
1900	44,633
1910	78,800
1920	138,276
1930	292,352
1940	384,514
1950	596,163
1960	938,219
1970	1,233,505
1980	1,595,138
1990	1,631,766
2000	1,953,631
2010	2,100,263

Give the scatterplot, LSLR, R^2 and the residual plot of this data. Can we say that the linear equation is the "best fit" for this data?

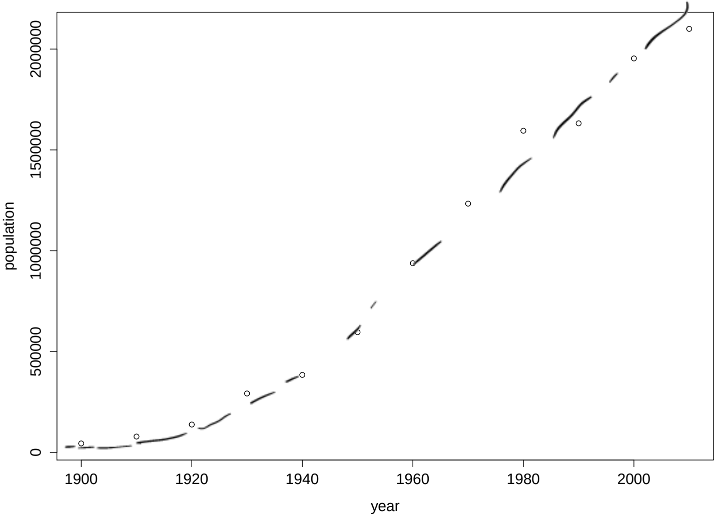
R code

```
year=c(1900,1910,1920,1930,1940,1950,1960,1970,1980,1990,2000,2010)
population=c(44633,78800,138276,292352,384514,596163,938219,1233505,
1595138,1631766,1953631,2100263)
plot(year,population)
pop.lm=lm(population~year)
pop.lm

Call:
lm(formula = population ~ year)
Coefficients:
(Intercept)         year
-39649140        20749
```

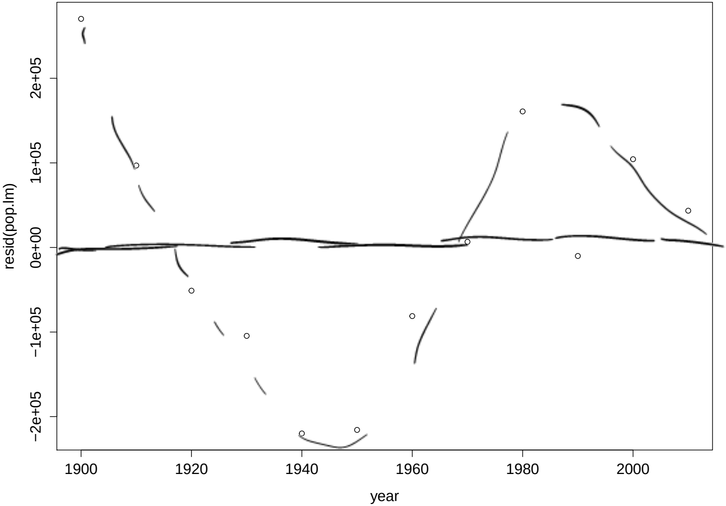
$\hat{y} = -39{,}649{,}104 + 20{,}749x$

```
cor(year,population)^2
[1] 0.9630567
plot(year,resid(pop.lm))
```

Scatterplot



Residual Plot



UNIVERSITY of HOUSTON DEPARTMENT OF MATHEMATICS

Results

- The coefficient of determination, R^2 is high.
- The scatter plot and residual plot shows a non-linear pattern.
- The least-square regression line might not be the "best fit" for this data.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Non-Linear Models

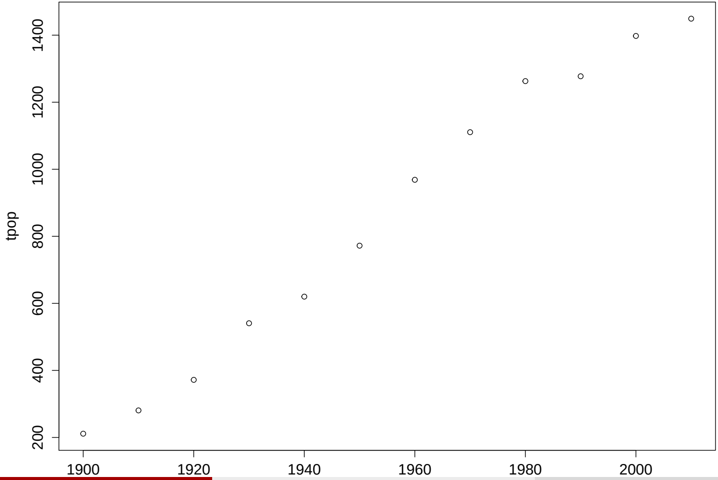
- Many times a scatterplot reveals a curved pattern instead of a linear pattern.
- We can transform the data by changing the scale of the measurement that was used when the data was collected.
- In order to find a good model we may need to transform our x value or our y value or both.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Transform Year

In our example it appears like the relationship is $y = x^2$. If we transform the points (x, y) into $(x, \sqrt{y})$ we get the scatterplot.

plot(x, sqrt(y))



R code

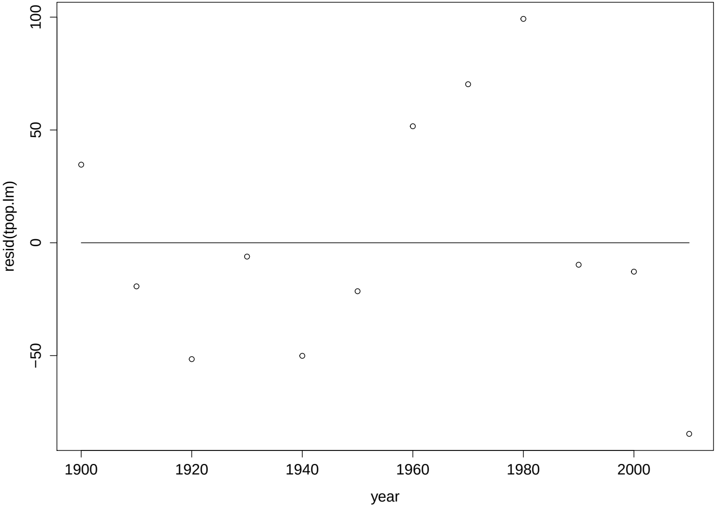
```
tpop=sqrt(population)
plot(year,tpop)
tpop.lm=lm(tpop~year)
tpop.lm

Call:
lm(formula = tpop ~ year)
Coefficients:
(Intercept)  year
-23267.19  12.34
```

$\Rightarrow \sqrt{\hat{y}} = -23267.19 + 12.34x$

```
cor(year,tpop)^2
[1] 0.9854597
plot(year,resid(tpop.lm))
lines(year,rep(0,length(year)))
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Residual Plot of Transformed Data



Calculating LSLR from Mosaic Date

Check mark the Mosaic Data in the programs list. The following is to create the scatterplot, LSLR, and the residual plot for predicting **temperature** in Celsius for water based on **time** in minutes that water was poured into a mug. Let X = time and Y = temp.

```
x = CoolingWater$time
y = CoolingWater$temp
plot(x,y)
cor(x,y)^2
[1] 0.778089
water.lm=lm(y~x)
water.lm

Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
64.2766          -0.2164

plot(x,resid(water.lm))
```

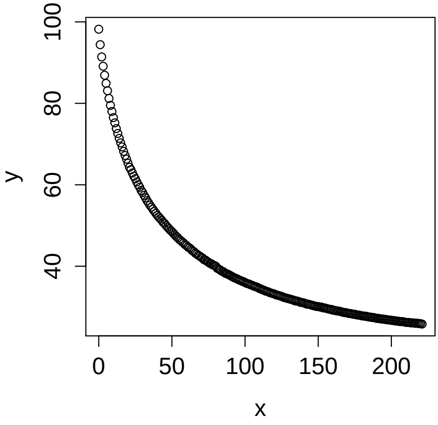
$R^2 = 77.8\%$

$\hat{y} = 64.2766 - 0.2164$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

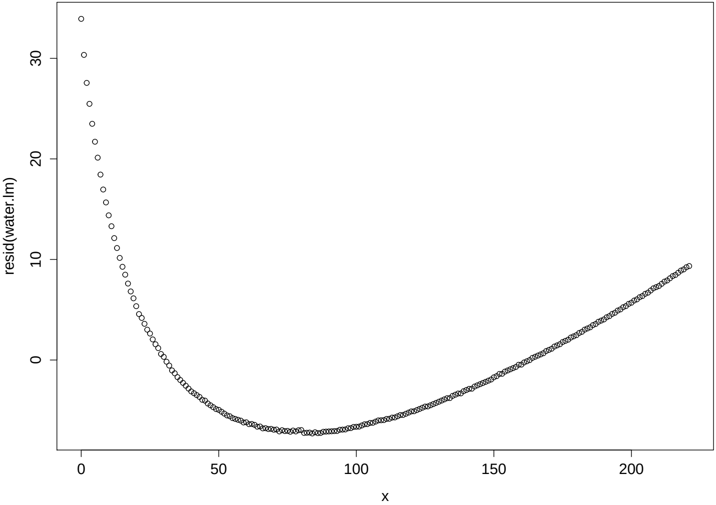
Scatterplot

Using this scatterplot, could we say that the LSLR is a good model to predict y?



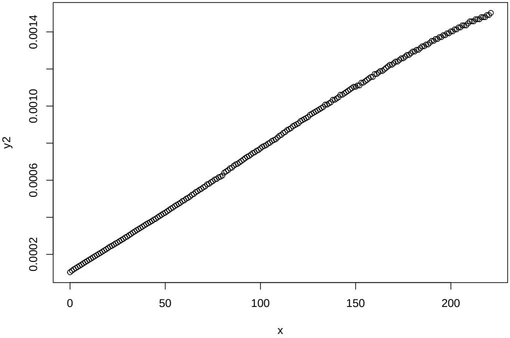
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Residual Plot



Transformation

A good transformation for this is $y = \frac{1}{\sqrt{x}}$. We will transform the points (x, y) to $(x, \frac{1}{y^2})$. plot(x, 1/y^2)



See your textbook for other examples of transformations.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

R code

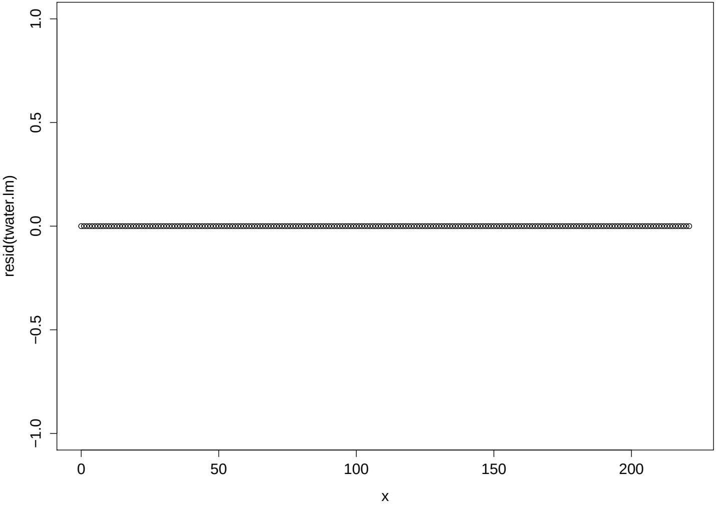
```
y2=1/y^2
twater.lm=lm(y2~x)
plot(x,resid(twater.lm),ylim=c(-1,1))
cor(x,y2)^2
[1] 0.9986831
```

$\Leftarrow R^2$

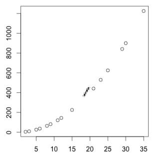
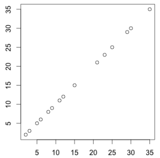
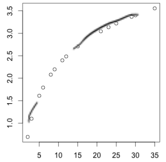
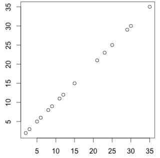
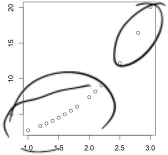
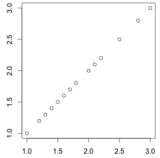
Residual Plot

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Residual Plot



Other Possible Transfers

Original Scatter-plot	Type of plot and Transformation	Modified Scatter-plot
	← looks like $y = x^2$ transform by changing (x, y) into $(x, \sqrt{y})$ →	
	← looks like $y = \log x$ transform by changing (x, y) into (x, e^y) →	
	← looks like $y = e^x$ transform by changing (x, y) into $(x, \log y)$ →	

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS