Section 5.5
Non-Linear Models

Many times a scatter-plot reveals a curved pattern instead of a linear pattern. When this is the case, the LSRL will not be an appropriate model for the data. Using functions such as the square root function or the logarithm function, 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.

Follow the guide below to know what to try when transforming data.

Original Data Scatter-plot looks like:	**Transform by changing:**
$y = x^2$ for (x, y)	(x, y) to $(x, \sqrt{y})$
$y = e^x$ for (x, y)	(x, y) to $(x, \log y)$
$y = \log x$ for (x, y)	(x, y) to (x, e^y)
$y = \frac{1}{x}$ for (x, y)	(x, y) to $(x, \frac{1}{y})$
$y = \frac{1}{x^2}$ for (x, y)	(x, y) to $(x, \frac{1}{\sqrt{y}})$

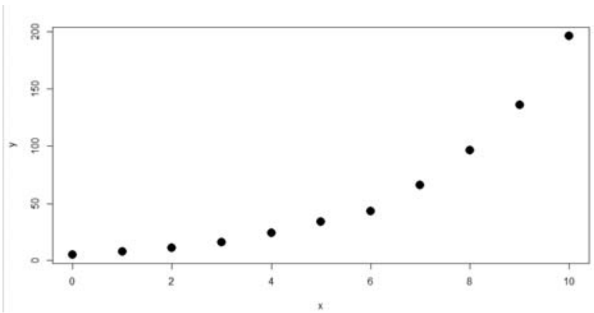
Then look at the residuals (and correlation) to determine if the model is a good one or not.

Example: The number of cell phone subscribers dramatically increased between the years 1990 and 2000. What kind of model would represent this data?

Year	Number of Subscribers (millions)
1990	5.3
1991	7.6
1992	11
1993	16
1994	24.1
1995	33.8
1996	43.4
1997	66.1
1998	95.8
1999	135.6
2000	196.2

Let’s begin by creating two lists, creating a scatter plot and finding its correlation.

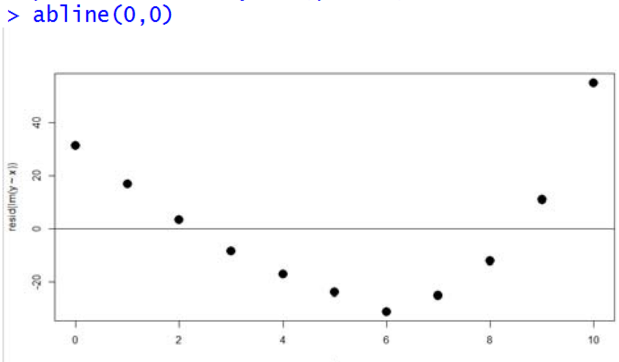
```
> x=c(0,1,2,3,4,5,6,7,8,9,10)
> y=c(5.3,7.6,11,16,24.1,33.8,43.4,66.1,95.8,135.6,196.2)
> plot(x,y,pch=16,cex=2)
> cor(x,y)
[1] 0.9019148
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -25.94        16.73
```



Is a linear model a good fit?

Now let's look at the residual plot.

```
> resid(lm(y~x))
         1          2          3          4          5
 31.236364  16.805455   3.474545  -8.256364 -16.887273
         6          7          8          9         10
-23.918182 -31.049091 -25.080000 -12.110909  10.958182
        11
 54.827273
> plot(x,resid(lm(y~x)))
> plot(x,resid(lm(y~x)),pch=16,cex=2)
> abline(0,0)
```



Since the oringial scatter plot looks quadratic or even exponential, let's first try transforming the data by assuming it's a quadratic.

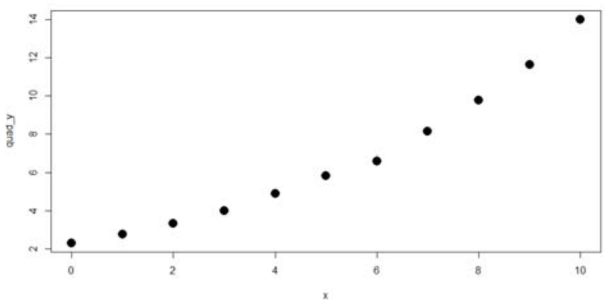
QUADRATIC TRANSFROMATION

Recall:

Original Data Scatter-plot looks like: $y = x^2$ for (x, y)

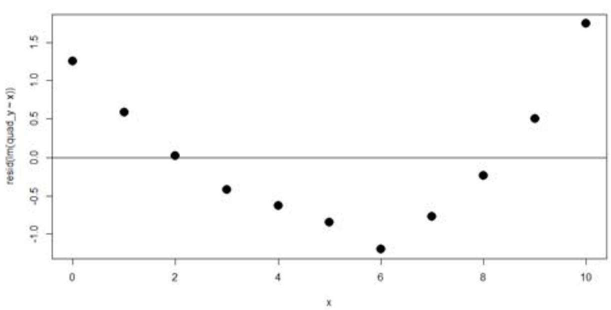
Transform by changing: (x, y) to $(x, \sqrt{y})$

```
> quad_y=sqrt(y)
> plot(x,quad_y,pch=16,cex=2)
> cor(x,quad_y)
[1] 0.9705223
> lm(quad_y~x)

Call:
lm(formula = quad_y ~ x)

Coefficients:
(Intercept)            x
      1.049        1.122
```



Notice how this model for the new data is better than before, not as curvy as the original.

```
> plot(x,resid(lm(quad_y~x)))
> plot(x,resid(lm(quad_y~x)),pch=16,cex=2)
> abline(0,0)
```



How does this look? Any better?

Now let's try transforming the data based on assuming the original data was exponential.

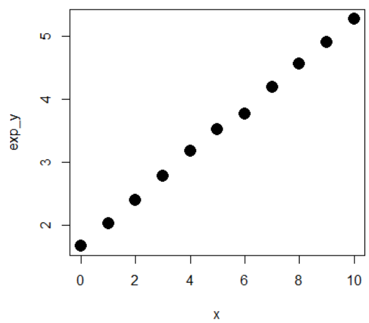
EXPONENTIAL TRANSFROMATION

Recall:

Original Data Scatter-plot looks like:
$y = e^x$ for (x, y)

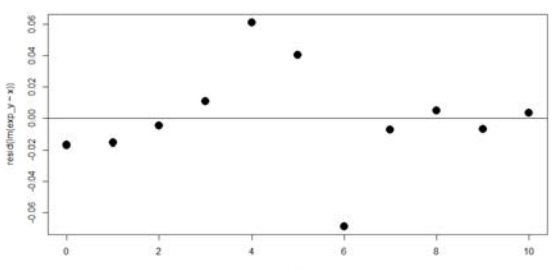
Transform by changing:
(x, y) to $(x, \log y)$

```
> exp_y=log(y)
> plot(x,exp_y,pch=16,cex=2)
> cor(x,exp_y)
[1] 0.999616
> lm(exp_y~x)

Call:
lm(formula = exp_y ~ x)

Coefficients:
(Intercept)            x
     1.6846       0.3591
```



How does this model for the new data here look? Better?

```
> plot(x,resid(lm(exp_y~x)),pch=16,cex=2)
> abline(0,0)
```



How do the residuals look?

So which model, quadratic or exponential, can we assume is the best fit model?

Lastly, give the final model for this problem: *Since the exponential was the best transformation use the LSRL for the exponenail and set it equal to log(y) and solve for y.*

Final Model

$$\log(y) = 0.3591x + 1.6846$$

$$e^{\log(y)} = e^{0.3591x+1.6846}$$

$$y = e^{0.3591x+1.6846}$$

R Commands:

```
>plot(x,y,pch=16,cex=2)
>curve(exp(0.3591*x+1.6846),add=TRUE)
```