STAT 420

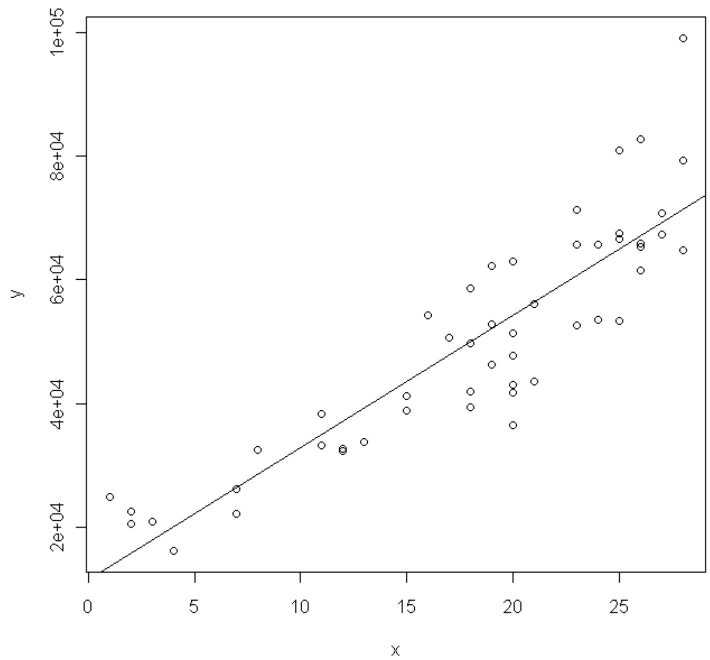
A useful **variance-stabilizing transformation** – $\ln y$.

1. The data in https://netfiles.uiuc.edu/stepanov/www/Initech.csv are the salaries, y, and years of experience, x, for a sample of 50 *Initech* employees.

```
> Initech
    x     y
1   7 26075
2  28 79370
3  23 65726
4  18 41983
5  19 62308
6  15 41154
7  24 53610
8  13 33697
9   2 22444
10  8 32562
11 20 43076
12 21 56000
13 18 58667
14  7 22210
15  2 20521
16 18 49727
17 11 33233
18 21 43628
19  4 16105
20 24 65644
21 20 63022
22 20 47780
23 15 38853
24 25 66537
25 25 67447
26 28 64785
27 26 61581
28 27 70678
29 20 51301
30 18 39346
31  1 24833
32 26 65929
33 20 41721
34 26 82641
35 28 99139
36 23 52624
37 17 50594
38 25 53272
39 26 65343
40 19 46216
41 16 54288
42  3 20844
43 12 32586
44 23 71235
45 20 36530
46 19 52745
47 27 67282
48 25 80931
49 12 32303
50 11 38371
```

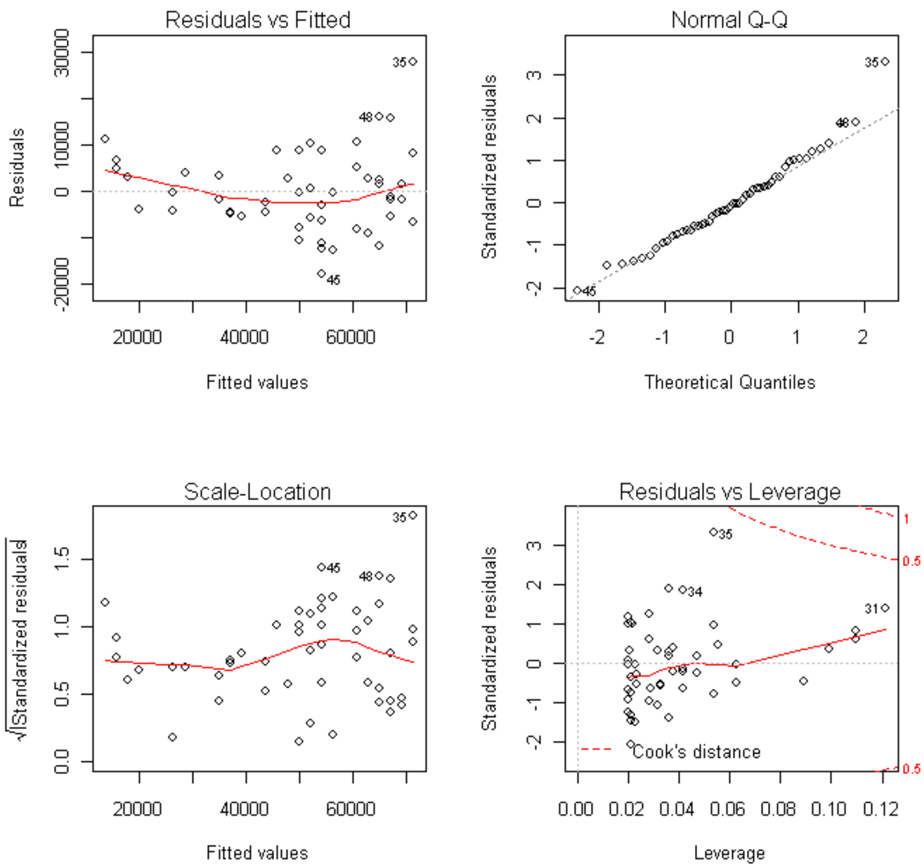
First, consider the first-order model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$.

```
> fit1 = lm(y ~ x)

> plot(x, y)
> abline(fit1$coefficients)
```



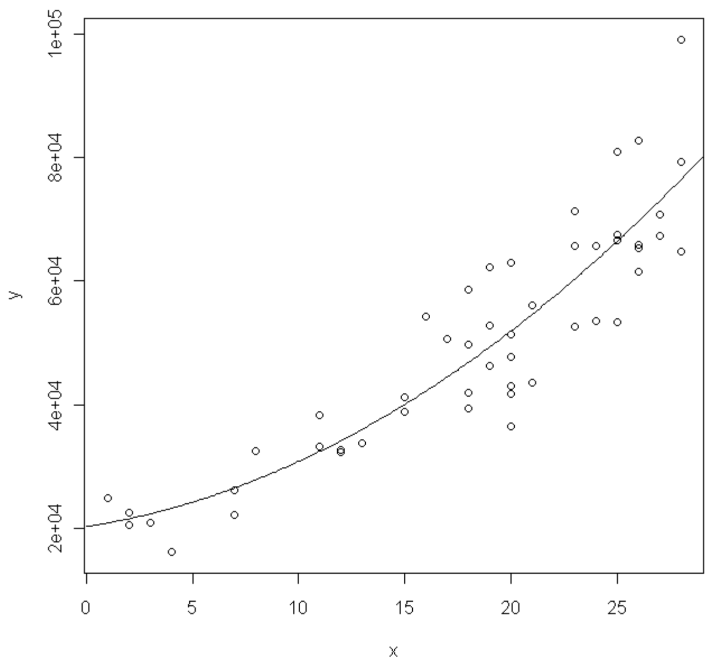

```
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-17665.6  -5497.7   -725.7   4667.3  27812.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11369.4     3160.2   3.598 0.000757 ***
x             2141.3      160.8  13.314  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8642 on 48 degrees of freedom
Multiple R-Squared: 0.7869,     Adjusted R-squared: 0.7825
F-statistic: 177.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(fit1)
```



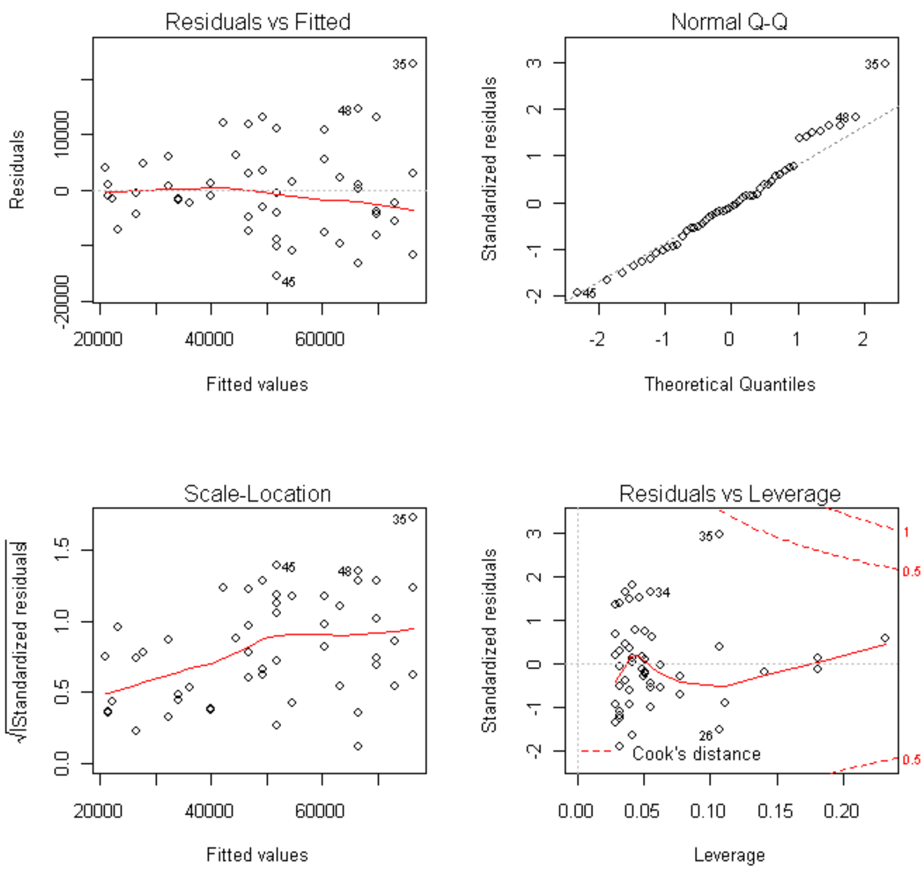
Consider the second-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$.

```
> fit2 = lm(y ~ x + I(x^2))

> plot(x, y)
> xx = seq(0,30,by=0.1)
> yy = fit2$coeff[1] + fit2$coeff[2]*xx + fit2$coeff[3]*xx^2
> lines(xx,yy)
```




```
> summary(fit2)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-15359.8  -4702.8   -783.3   3872.9  22718.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20241.98    4422.55   4.577 3.46e-05 ***
x             522.38     616.67   0.847  0.40124
I(x^2)         53.00      19.57   2.708  0.00941 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8123 on 47 degrees of freedom
Multiple R-Squared: 0.8157,     Adjusted R-squared: 0.8078
F-statistic:   104 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
> plot(fit2)
```

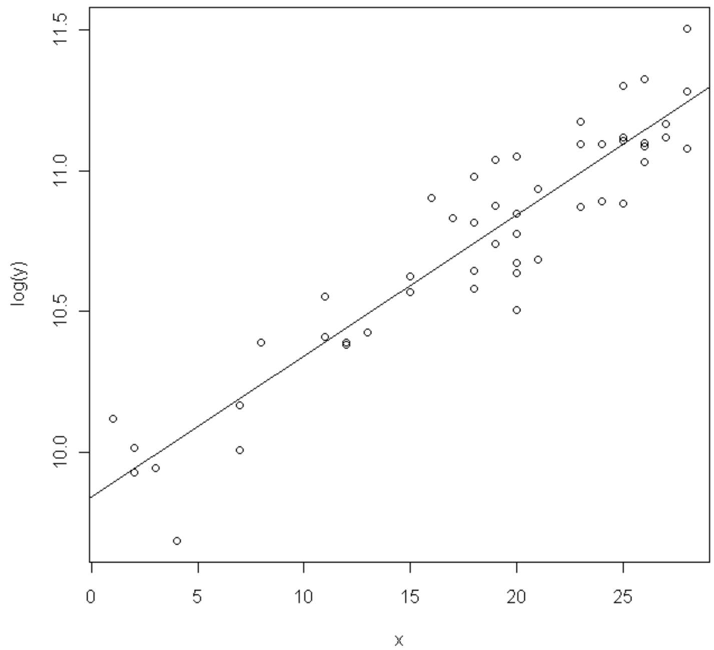


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

A useful **variance-stabilizing transformation** – $\ln y$.

Consider the model $\ln Y = \beta_0 + \beta_1 x_1 + \varepsilon$.

```
> fit3 = lm(log(y) ~ x)
> plot(x, log(y))
> abline(fit3$coefficients)
```




```
> summary(fit3)

Call:
lm(formula = lny ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35435 -0.09045 -0.01726  0.09740  0.26357 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.841325   0.056355  174.63   <2e-16 ***
x           0.049978   0.002868   17.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.1541 on 48 degrees of freedom
Multiple R-Squared: 0.8635,     Adjusted R-squared: 0.8607 
F-statistic: 303.6 on 1 and 48 DF,  p-value: < 2.2e-16 
```

The fitted regression function is

$$\ln(y) = 9.841325 + 0.049978 * x.$$

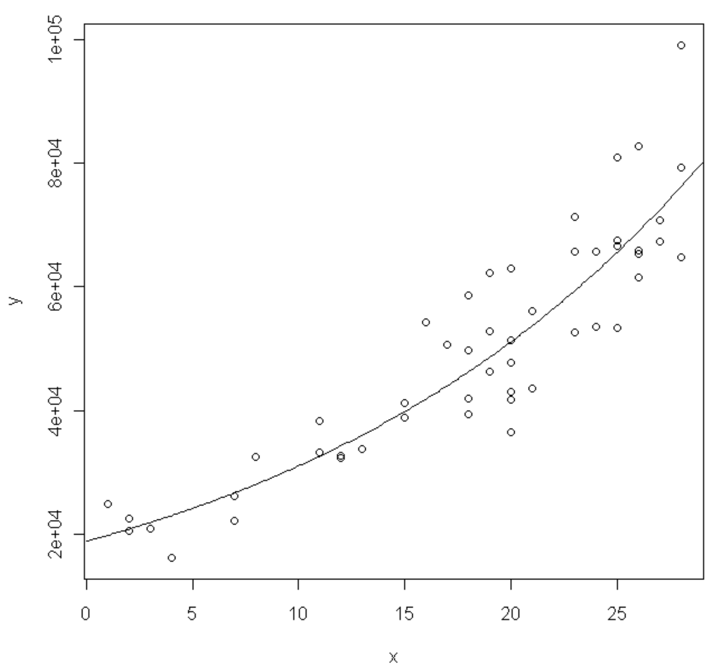
That is,

$$\text{wage} = e^{9.841325} \cdot e^{0.049978 * x}.$$

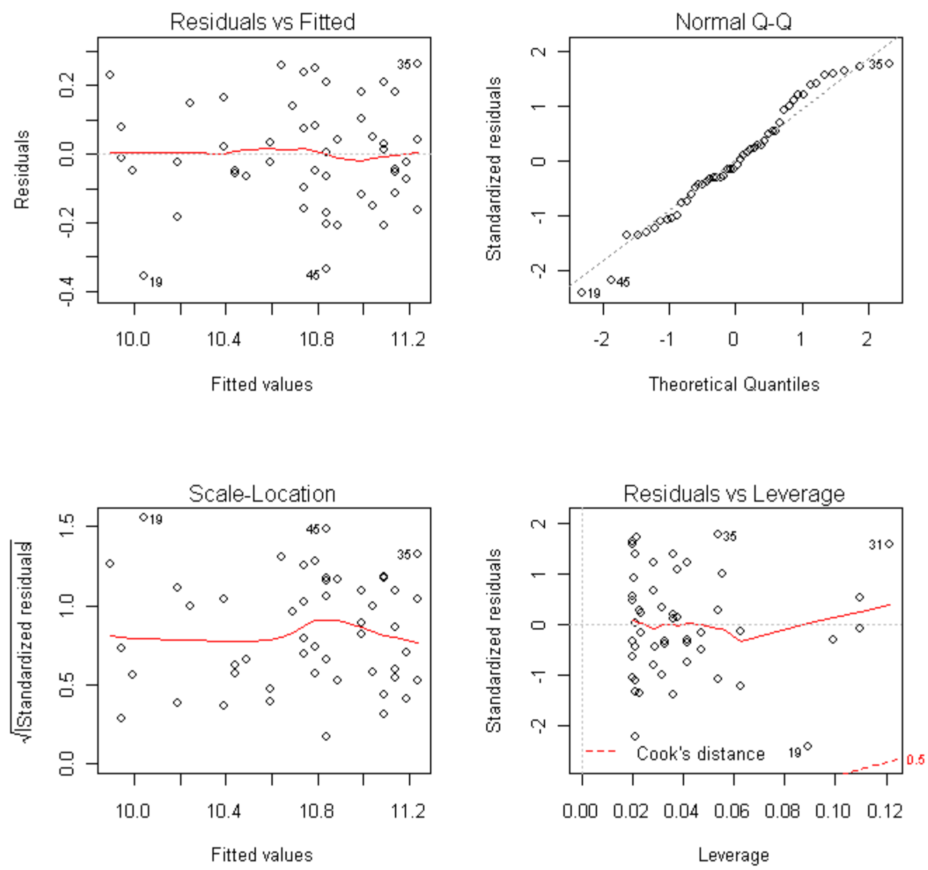
We would expect weekly wages to increase $e^{0.049978} = 1.051248$ times ("on average")

[that is, by 5.125%] for every 1-year increase of years of experience.

```
> plot(x, y)
> xx = seq(0,30,by=0.1)
> yy2 = exp(fit3$coeff[1] + fit3$coeff[2]*xx)
> lines(xx,yy2)
```



```
> plot(fit3)
```



Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

2. An experiment is conducted for the purpose of studying the effect of temperature on the life-length of an electrical insulation. Specimens of the insulation were tested under fixed temperatures, and their times to failure recorded.

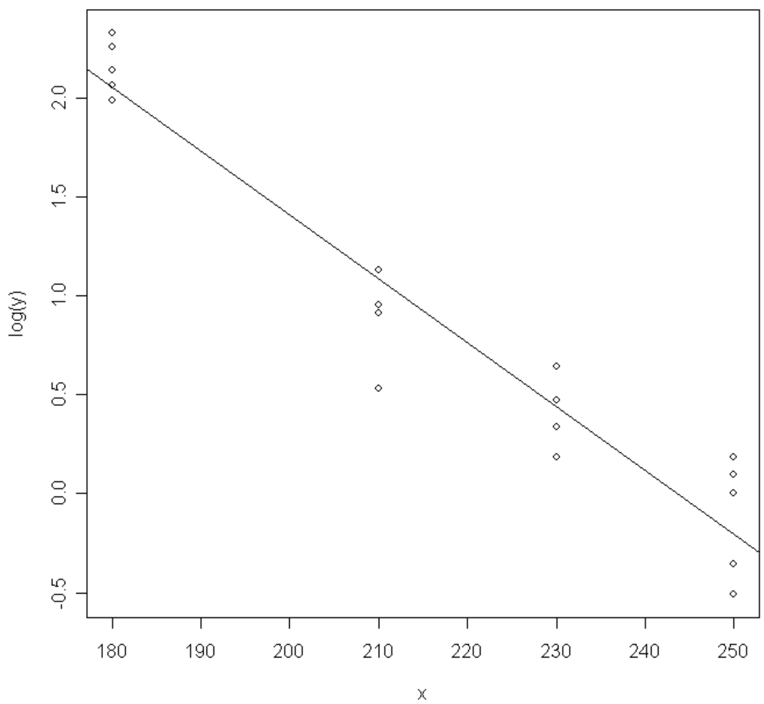
Temperature x (°C)	Failure Time y (thousand hours)
180	7.3, 7.9, 8.5, 9.6, 10.3
210	1.7, 2.5, 2.6, 3.1
230	1.2, 1.4, 1.6, 1.9
250	0.6, 0.7, 1.0, 1.1, 1.2

```
> x = c(rep(180,5), rep(210,4), rep(230,4), rep(250,5))
> y = c(7.3,7.9,8.5,9.6,10.3, 1.7,2.5,2.6,3.1, 1.2,1.4,1.6,1.9,
0.6,0.7,1.0,1.1,1.2)

> plot(x,y)
```

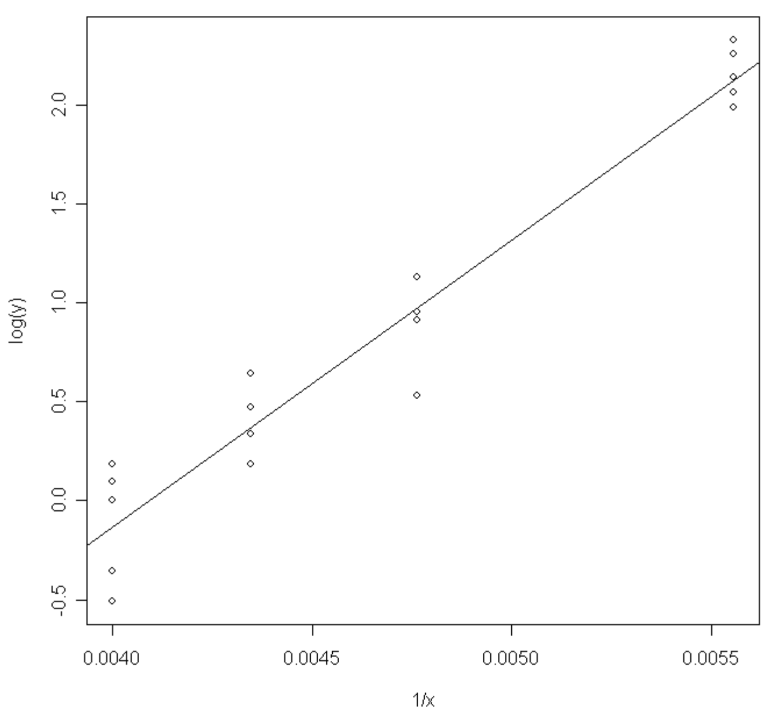
Relationship is non-linear, spread of y values is different for different x values.

$\Rightarrow$ Try $y' = \ln y$.

```
> plot(x,log(y))
> abline(lm(log(y)~x)$coefficients)
```




```
> plot(1/x,log(y))
> abline(lm(log(y)~I(1/x))$coefficients)
```



$\ln y$ vs. $\dfrac{1}{x}$ seems to be a better fit.

a) Fit a straight-line regression to the transformed data $x' = \dfrac{1}{x}$ and $y' = \ln y$

```
> fit = lm(log(y)~I(1/x))
> summary(fit)

Call:
lm(formula = log(y) ~ I(1/x))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.443853 -0.120753 -0.003639  0.151037  0.315577 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -5.9489     0.4076  -14.59 1.15e-10 ***
I(1/x)      1453.9049    86.4045   16.83 1.35e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2222 on 16 degrees of freedom
Multiple R-squared: 0.9465,     Adjusted R-squared: 0.9432 
F-statistic: 283.1 on 1 and 16 DF,  p-value: 1.348e-11
```

b) Is there strong evidence that an increase in temperature reduces the life of the insulation?

Yes. β_1 is very significant. $\hat{\beta}_1$ is positive. The evidence is strong that $\ln y$ depends on ${}^1\!/_x$. The expected value of $\ln(\text{Failure Time})$ decreases as ${}^1\!/_{\text{Temperature}}$ decreases. Thus increasing the temperature reduces the life of the insulation.

c) Predict the Failure Time y when $x = 200$ C.

$x = 200 \quad \Rightarrow \quad x' = 0.005$

$\Rightarrow \quad \hat{y}' = -5.9489 + 1453.9049 \cdot 0.005 = 1.3206$

$\Rightarrow \quad \hat{y} = e^{1.3206} = \mathbf{3.746}$.

```
> predict(fit,data.frame(x=200))
[1] 1.320650
> exp(1.320650)
[1] 3.745855
```

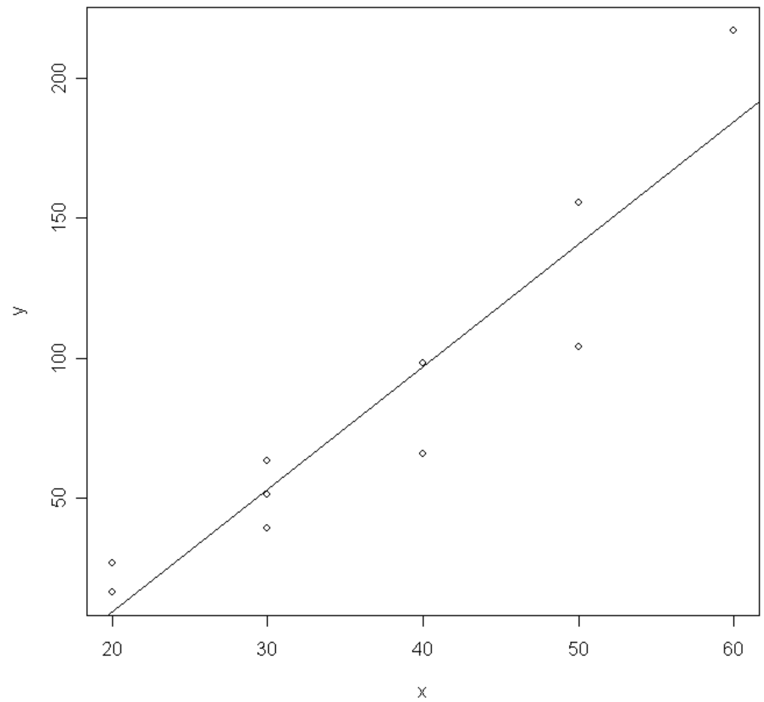
```
> predict.lm(fit,data.frame(x=200),interval=c("prediction"))
          fit       lwr      upr
[1,] 1.320650 0.8331375 1.808162
> exp(0.8331375)
[1] 2.300525
> exp(1.808162)
[1] 6.099227
```

95% prediction interval: $(2.3, 6.1)$

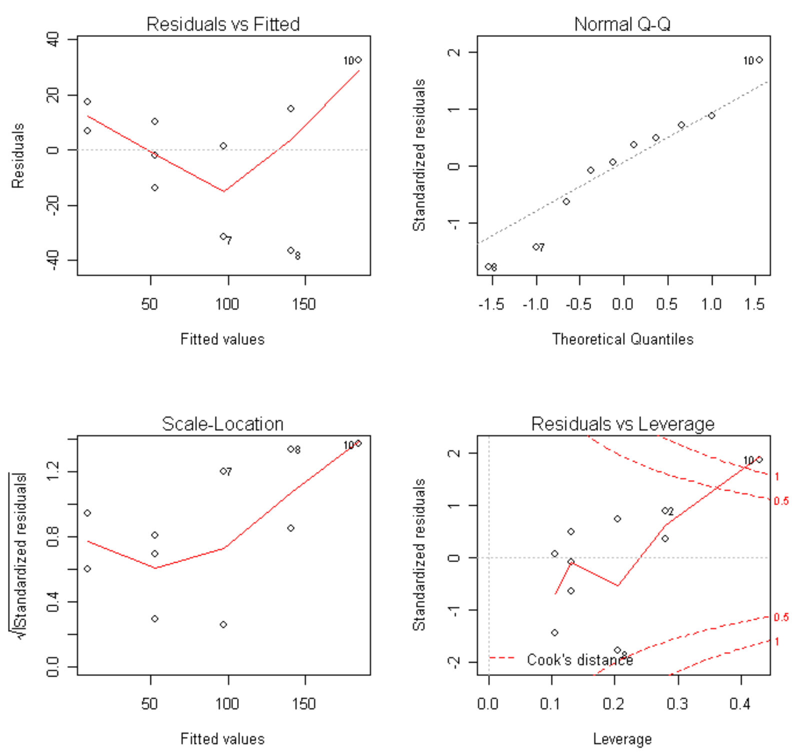
3. To determine the maximum stopping ability of cars when their brakes are fully applied, 10 cars are driven each at a specified speed and the distance each requires to come to a complete stop is measured. The various initial speeds selected for each of the 10 cars and the stopping distances recorded are given in the table on the right. Use transformations to find a good model for predicting the stopping distance from the initial speed and use it to predict the stopping distance y if the initial speed is $x = 55$ mph.

Initial Speed x (mph)	Stopping Distance y (ft)
20	16.3
20	26.7
30	39.2
30	63.5
30	51.3
40	98.4
40	65.7
50	104.1
50	155.6
60	217.2

```
> x = c(20,20,30,30,30,40,40,50,50,60)
> y = c(16.3,26.7,39.2,63.5,51.3,98.4,65.7,104.1,155.6,217.2)

> fit1 = lm(y ~ x)

> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-36.666 -10.901    4.224   13.719   32.614

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -78.3335    22.5030  -3.481   0.0083 **
x             4.3820     0.5753   7.617  6.2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.08 on 8 degrees of freedom
Multiple R-Squared: 0.8788,     Adjusted R-squared: 0.8637
F-statistic: 58.02 on 1 and 8 DF,  p-value: 6.206e-05


> plot(x,y)
> abline(fit1$coefficients)
```



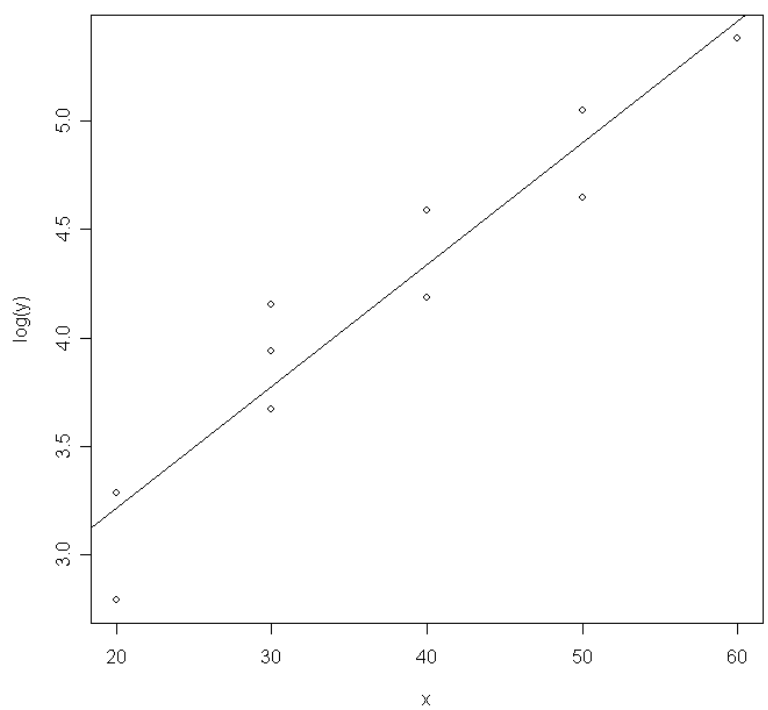
```
> par(mfrow=c(2,2))
> plot(fit1)
```



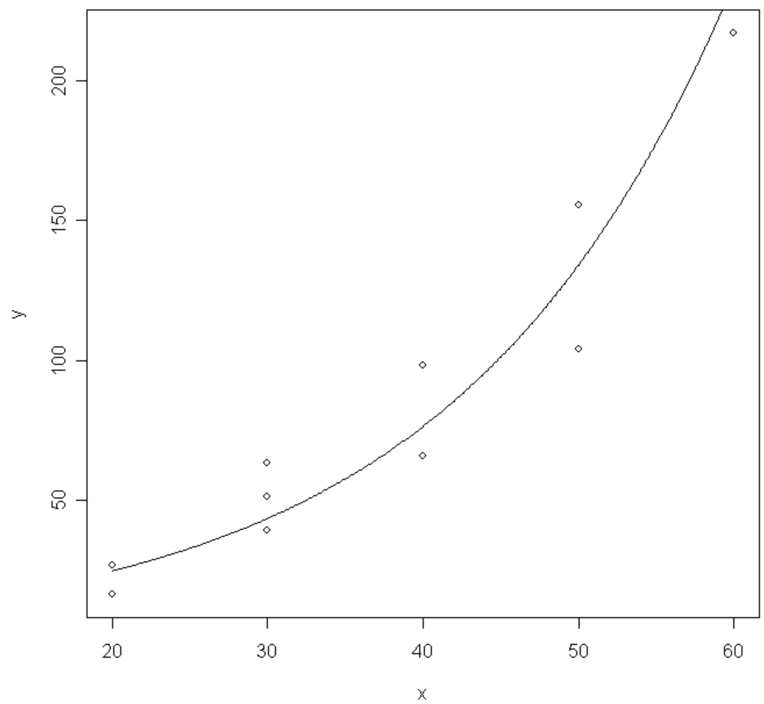
If the relationship between x and y is exponential, $y = A\, e^{b\,x}$,

then $\ln y = \ln A + b\,x$, or $z = a + b\,x$, where $z = \ln y$, $a = \ln A$.

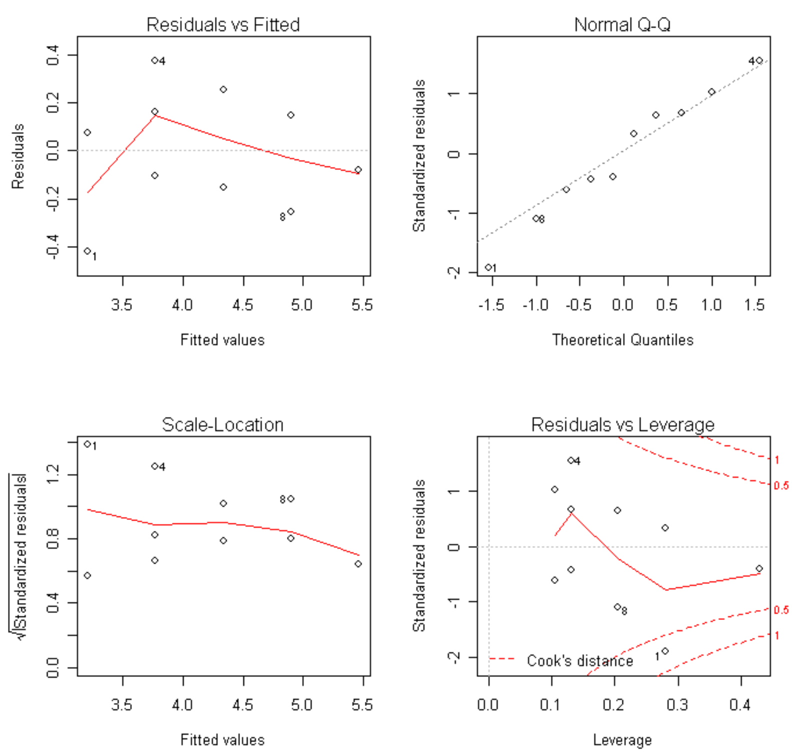
```
> fit2 = lm(log(y) ~ x)
> summary(fit2)

Call:
lm(formula = log(y) ~ x)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.421385 -0.140194 -0.003966  0.159422  0.376411 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.088393   0.253648   8.233 3.55e-05 ***
x           0.056208   0.006485   8.668 2.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2602 on 8 degrees of freedom
Multiple R-squared: 0.9038,     Adjusted R-squared: 0.8917 
F-statistic: 75.13 on 1 and 8 DF,  p-value: 2.441e-05 

> plot(x,log(y))
> abline(fit2$coefficients)
```



```
> plot(x,y)
> xx = seq(20,60,by=0.1)
> yy2 = exp(fit2$coeff[1]+fit2$coeff[2]*xx)
> lines(xx,yy2)
```



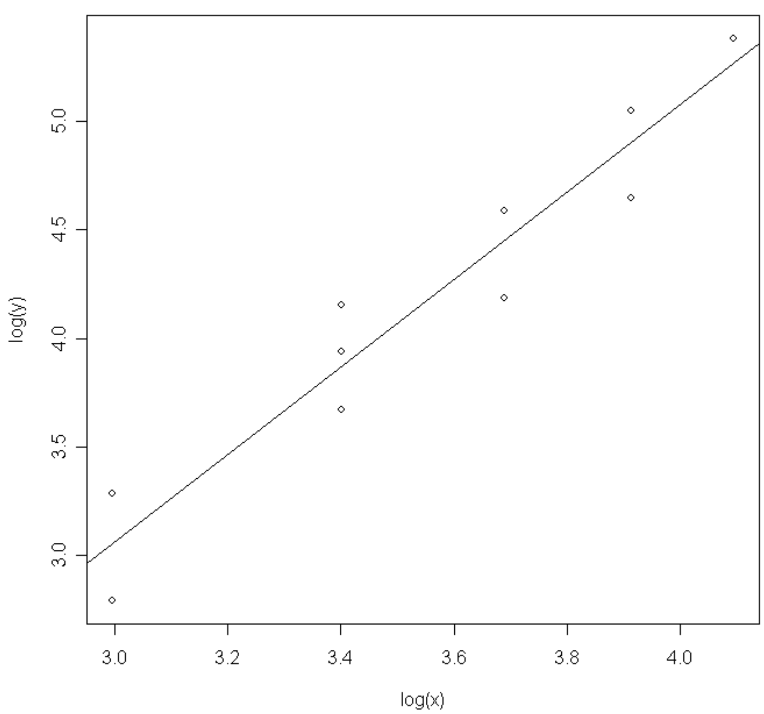

```
> plot(fit2)
```



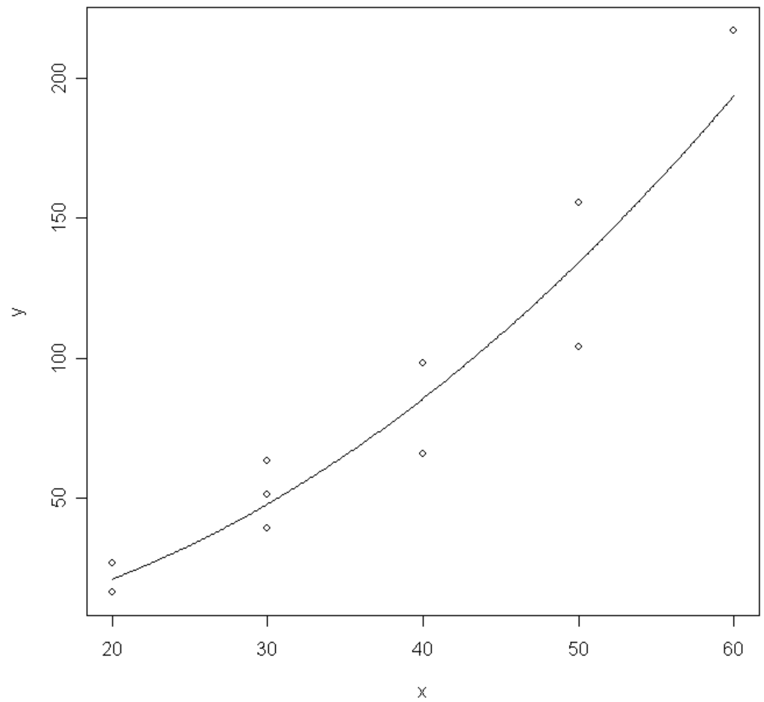
If the relationship between x and y is power function, $y = A\,x^{b}$,

then $\ln y = \ln A + b \ln x$, or $z = a + b\,w$, where $z = \ln y$, $w = \ln x$, $a = \ln A$.

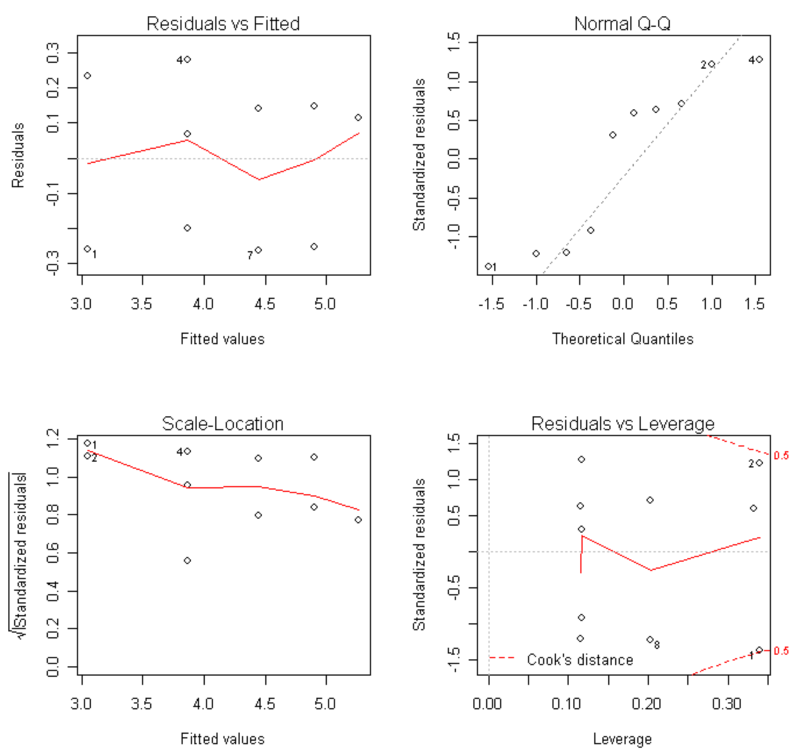
```
> fit3 = lm(log(y) ~ log(x))
> summary(fit3)

Call:
lm(formula = log(y) ~ log(x))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26472 -0.24103  0.09071  0.14553  0.28115 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.9866     0.7360  -4.058  0.00364 ** 
log(x)        2.0159     0.2063   9.771 1.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2332 on 8 degrees of freedom
Multiple R-squared: 0.9227,     Adjusted R-squared: 0.913 
F-statistic: 95.47 on 1 and 8 DF,  p-value: 1.009e-05

> plot(log(x),log(y))
> abline(fit3$coefficients)
```



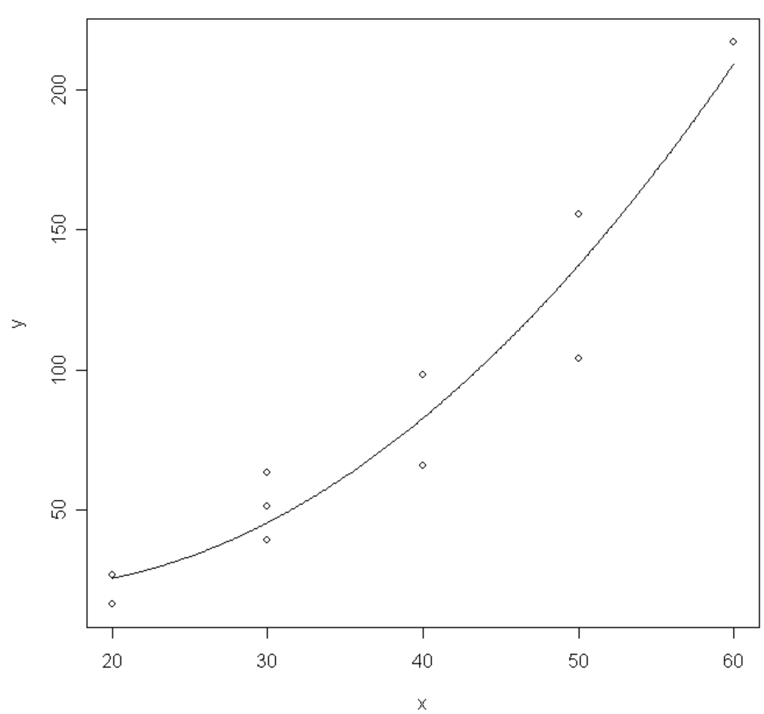
```
> plot(x,y)
> yy3 = exp(fit3$coeff[1]+fit3$coeff[2]*log(xx))
> lines(xx,yy3)
```



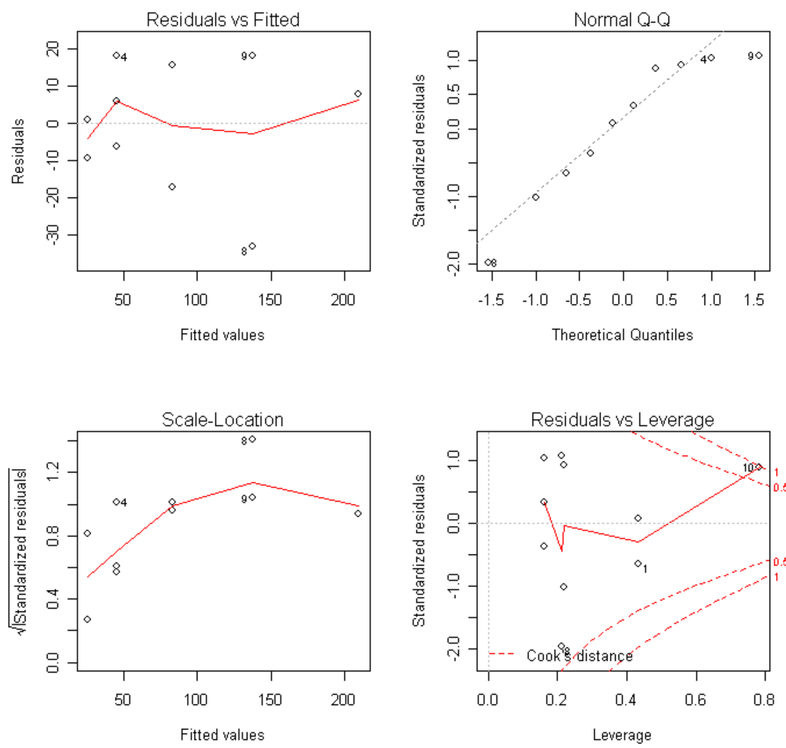

```
> plot(fit3)
```



```
> fit4 = lm(y ~ x + I(x^2))
> summary(fit4)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-33.326  -8.632   3.367  13.634  18.174 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 37.90064   56.01200   0.677   0.5204  
x           -2.34538    3.09621  -0.757   0.4735  
I(x^2)       0.08672    0.03944   2.199   0.0639 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 18.98 on 7 degrees of freedom
Multiple R-squared: 0.9283,     Adjusted R-squared: 0.9078 
F-statistic: 45.33 on 2 and 7 DF,  p-value: 9.861e-05 

> plot(x,y)
> yy4 = fit4$coeff[1]+fit4$coeff[2]*xx+fit4$coeff[3]*xx^2
> lines(xx,yy4)
```

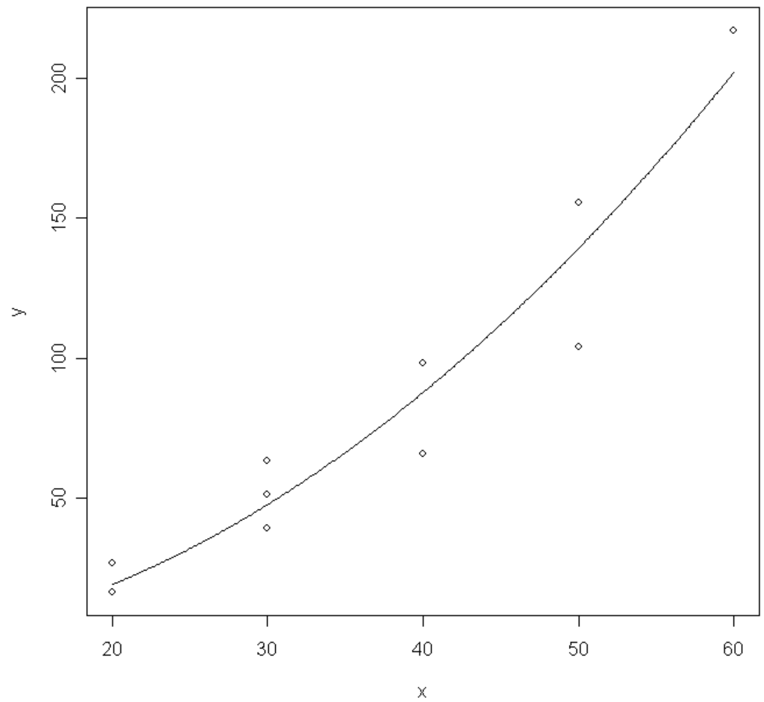


```
> plot(fit4)
```

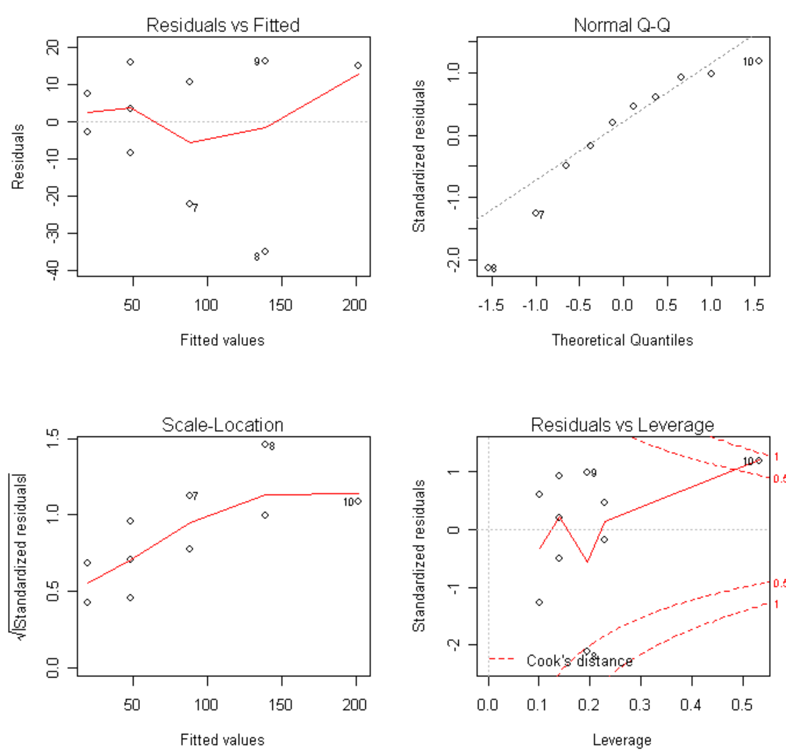



```
> fit5 = lm(y ~ I(x^2))
> summary(fit5)

Call:
lm(formula = y ~ I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-35.175  -7.146   5.528  13.910  16.325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.701935  10.703929  -0.346    0.738
I(x^2)       0.057191   0.005863   9.754 1.02e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.47 on 8 degrees of freedom
Multiple R-squared: 0.9224,     Adjusted R-squared: 0.9127
F-statistic: 95.15 on 1 and 8 DF,  p-value: 1.022e-05
```

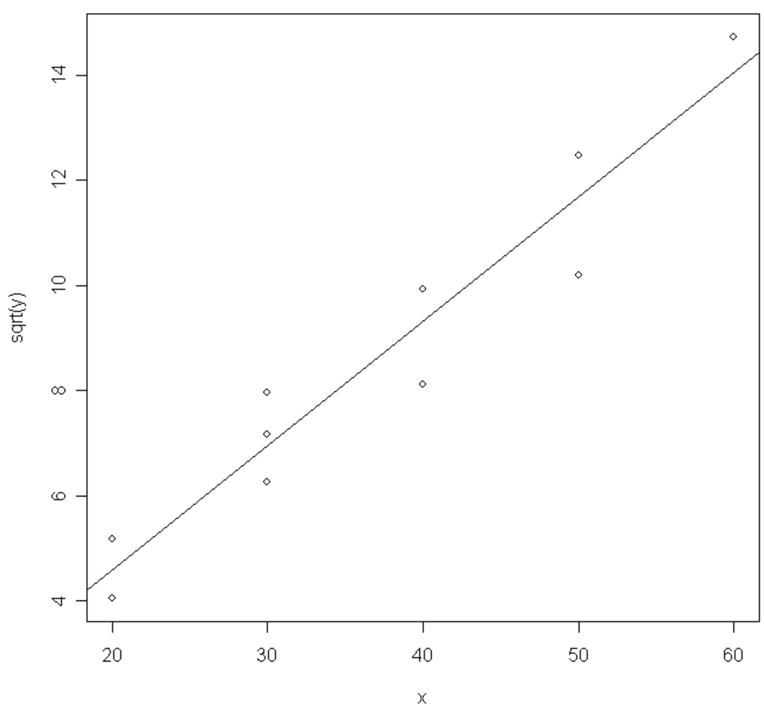
```
> plot(x,y)
> yy5 = fit5$coeff[1]+fit5$coeff[2]*xx^2
> lines(xx,yy5)
```



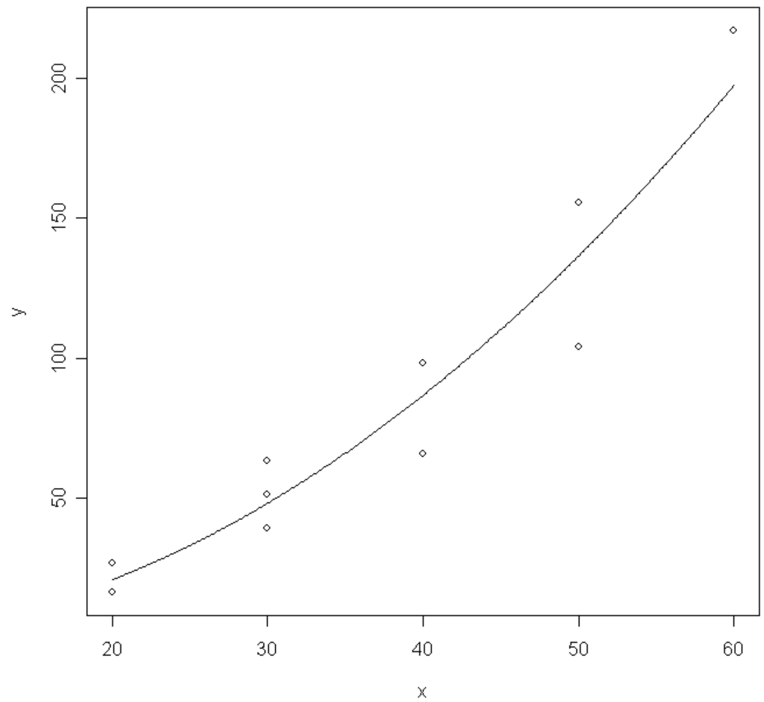

```
> plot(fit5)
```



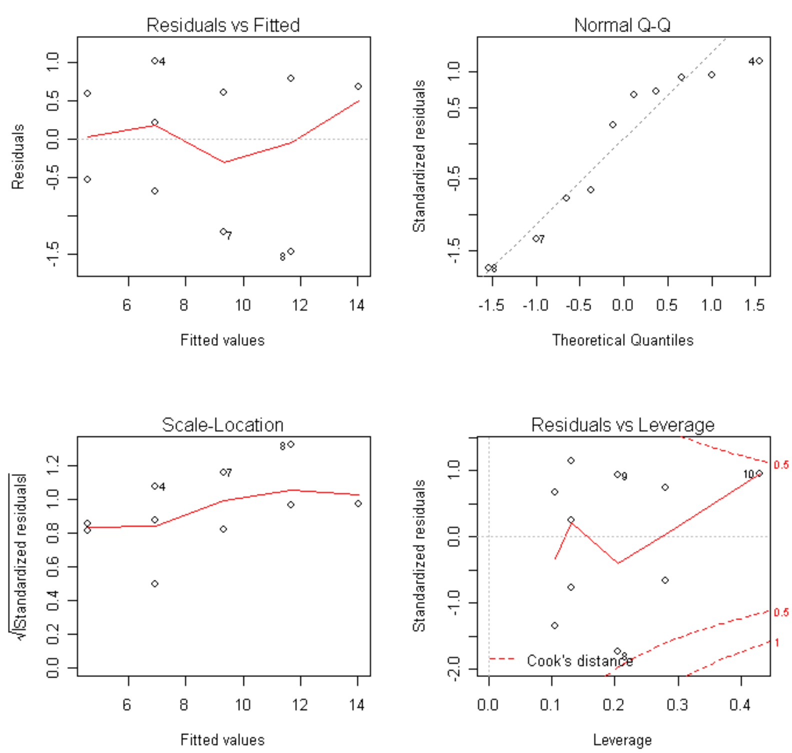
```
> fit6 = lm(sqrt(y) ~ x)
> summary(fit6)

Call:
lm(formula = sqrt(y) ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4821 -0.6468  0.4055  0.6630  1.0243 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.16653    0.93225  -0.179    0.863    
x            0.23703    0.02383   9.945 8.84e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9563 on 8 degrees of freedom
Multiple R-squared: 0.9252,     Adjusted R-squared: 0.9158 
F-statistic: 98.91 on 1 and 8 DF,  p-value: 8.843e-06


> plot(x,sqrt(y))
> abline(fit6$coefficients)
```



```
> plot(x,y)
> yy6 = (fit6$coeff[1]+fit6$coeff[2]*xx)^2
> lines(xx,yy6)
```




```
> plot(fit6)
```



`fit3` and `fit6` seem to be the best.

```
> new = data.frame(x=55)
```

```
> predict(fit3,new)
[1] 5.09179
> exp(5.09179)
[1] 162.6808
```

```
> predict(fit6,new)
[1] 12.87022
> 12.87022^2
[1] 165.6426
```

STAT 420

Box-Cox method:

Transformation $$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

Recall Example 3 from Examples for 10/11/12 (part 2):

3. To determine the maximum stopping ability of cars when their brakes are fully applied, 10 cars are driven each at a specified speed and the distance each requires to come to a complete stop is measured. The various initial speeds selected for each of the 10 cars and the stopping distances recorded are given in the table on the right. Use transformations to find a good model for predicting the stopping distance from the initial speed and use it to predict the stopping distance y if the initial speed is $x = 55$ mph.

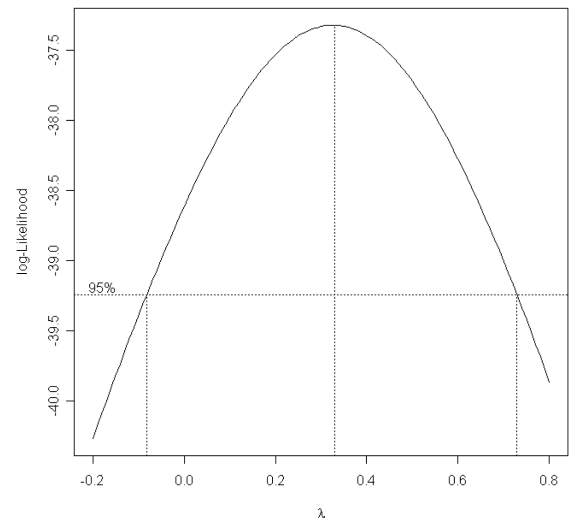
Initial Speed x (mph)	Stopping Distance y (ft)
20	16.3
20	26.7
30	39.2
30	63.5
30	51.3
40	98.4
40	65.7
50	104.1
50	155.6
60	217.2

```
> x = c(20,20,30,30,30,40,40,50,50,60)
> y = c(16.3,26.7,39.2,63.5,51.3,98.4,65.7,104.1,155.6,217.2)

> fit = lm(y ~ x)

> library(MASS)

> boxcox(fit,plotit=T)
```

```
> boxcox(fit,plotit=T,lambda=seq(-0.2,0.8,by=0.01))
```

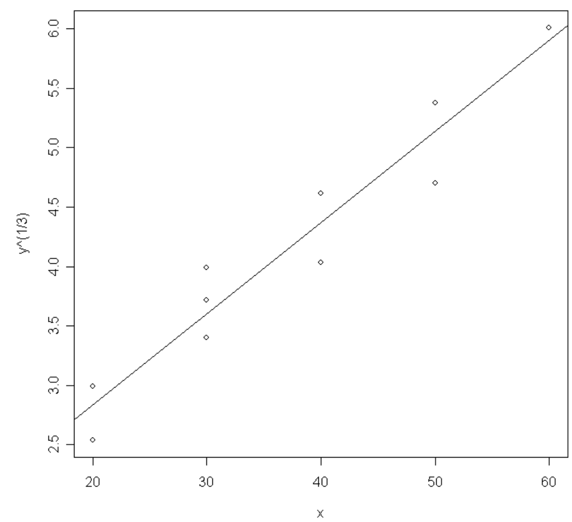
$\ln y$ ($\lambda = 0$) and $\sqrt{y}$ ($\lambda = 0.5$) are acceptable.

$\lambda \approx {}^{1}\!/_{3}$ seems to give the best transformation of the response variable.

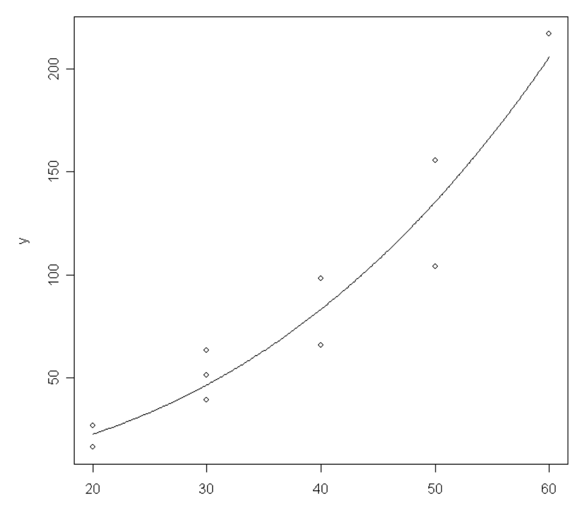



```
> fit2 = lm(y^(1/3) ~ x)
```

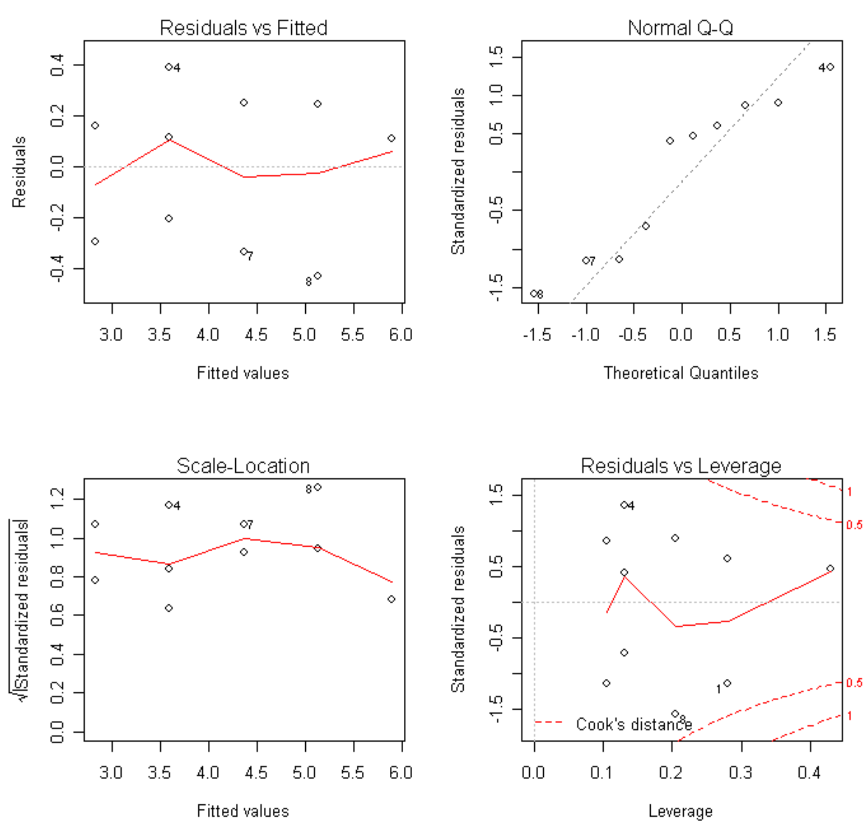
```
> plot(x,y^(1/3))
> abline(fit2$coefficients)
```




```
> plot(x,y)
> xx = seq(20,60,by=0.1)
> yy = (fit2$coeff[1]+fit2$coeff[2]*xx)^3
> lines(xx,yy)
```



```
> summary(fit2)

Call:
lm(formula = y^(1/3) ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4315 -0.2727  0.1117  0.2215  0.3900 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.295415   0.298150   4.345  0.00246 ** 
x           0.076806   0.007622  10.076 8.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.3058 on 8 degrees of freedom
Multiple R-squared: 0.927,     Adjusted R-squared: 0.9178 
F-statistic: 101.5 on 1 and 8 DF,  p-value: 8.019e-06 


> par(mfrow=c(2,2))
> plot(fit2)
```



- -

4. The data in the file were collected in a study of the effect of dissolved sulfur on the surface tension of liquid copper（Baes, C. and Kellogg, H. (1953). Effect of dissolved sulphur on the surface tension of liquid copper. *J. Metals*, 5, 643-648）.

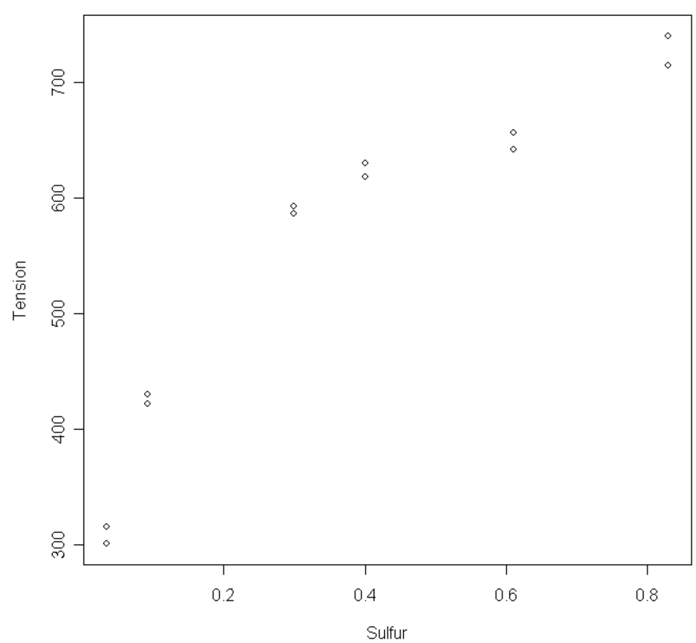
http://www.stat.umn.edu/alr/data/baeskel.txt

This data frame contains the following columns:

Sulfur	Weight percent sulfur
Tension	Decrease in surface tension, dynes/cm

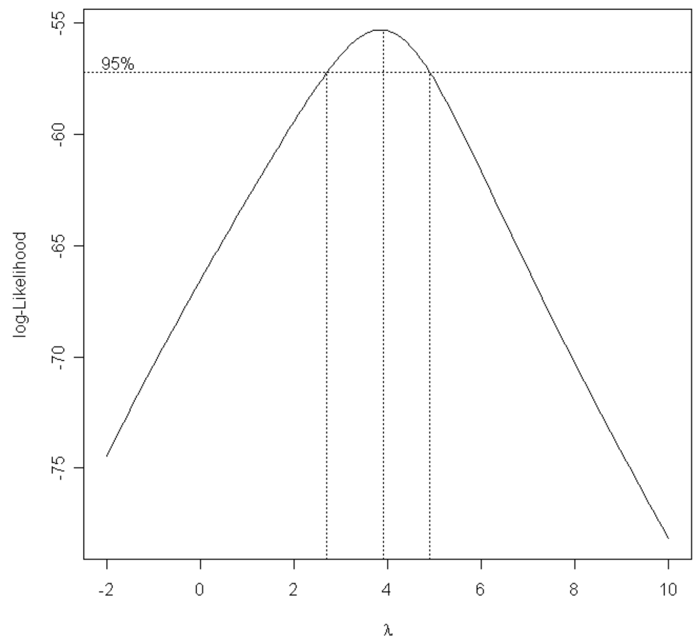
```
> tension = read.table("http://www.stat.umn.edu/alr/data/baeskel.txt",
header=T)
> attach(tension)
> tension
   Sulfur Tension
1   0.034     301
2   0.034     316
3   0.093     430
4   0.093     422
5   0.300     593
6   0.300     586
7   0.400     630
8   0.400     618
9   0.610     656
10  0.610     642
11  0.830     740
12  0.830     714
```

```
> plot(Sulfur,Tension)
```

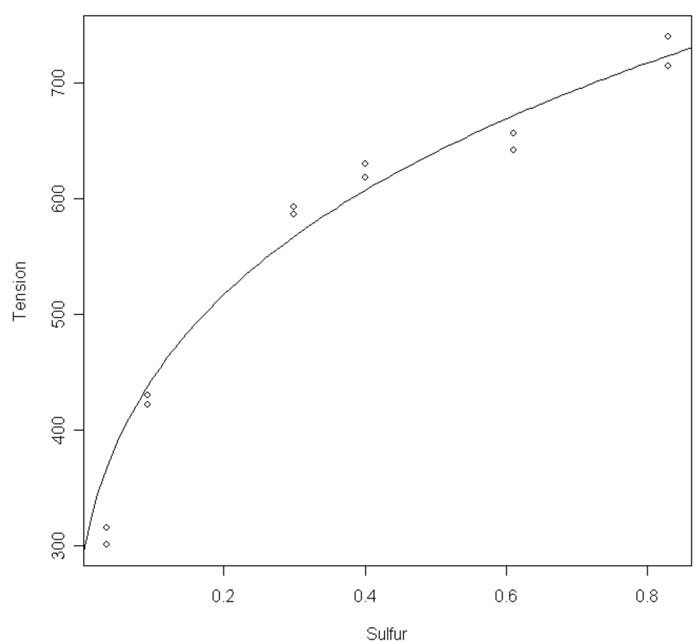


```
> fit = lm(Tension ~ Sulfur)

> boxcox(fit,plotit=T,lambda=seq(-2,10,by=0.1))
```

$\lambda \approx 4$ seems to give the best transformation of the response variable.




```
> fit2 = lm(Tension^4 ~ Sulfur)
> plot(Sulfur,Tension)
> x = seq(0,0.9,by=0.01)
> y = (fit2$coeff[1]+fit2$coeff[2]*x)^0.25
> lines(x,y)
```



A better fit:

```
> fit3 = lm(Tension ~ log(Sulfur))
> y2 = fit3$coeff[1]+fit3$coeff[2]*log(x)
> lines(x,y2,col=2)
```