Chapter 11

Data Transformations

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

This work is a derivative of 'Regression Methods' by Iain Pardoe, Laura Simon and Derek Young, used under CC BY-NC.

Contents

1. Transformaing Predictor

1. Transforming Predictors

- Transforming the x values is appropriate when non-linearity is the only problem — the independence, normality and equal variance conditions are met.

1. Transforming Predictors

- While some assumptions may appear to hold prior to applying a transformation, they may no longer hold once a transformation is applied. In other words, using transformations is part of an iterative process where all the linear regression assumptions are re-checked after each iteration.

1. Transforming Predictors

- Although we're focussing on a simple linear regression model here, the essential ideas apply more generally to multiple linear regression models too. We can consider transforming any of the predictors by examining scatterplots of the residuals versus each predictor in turn.

2. Data: Word Recall

- Data: Word Recall
  - Data from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items. The subjects were then asked to recall the items at various times up to a week later.
  - y (prop): proportion of items correctly recalled
  - x (time): time in minutes

3. Data Load

```
> wordrecall <- read.table("wordrecall.txt", header=T)
> attach(wordrecall)
> head(wordrecall)
  time prop
1    1 0.84
2    5 0.71
3   15 0.61
4   30 0.56
5   60 0.54
6  120 0.47
```

4. Model Creation

```
model.1 <- lm(prop ~ time)
summary(model.1)
```

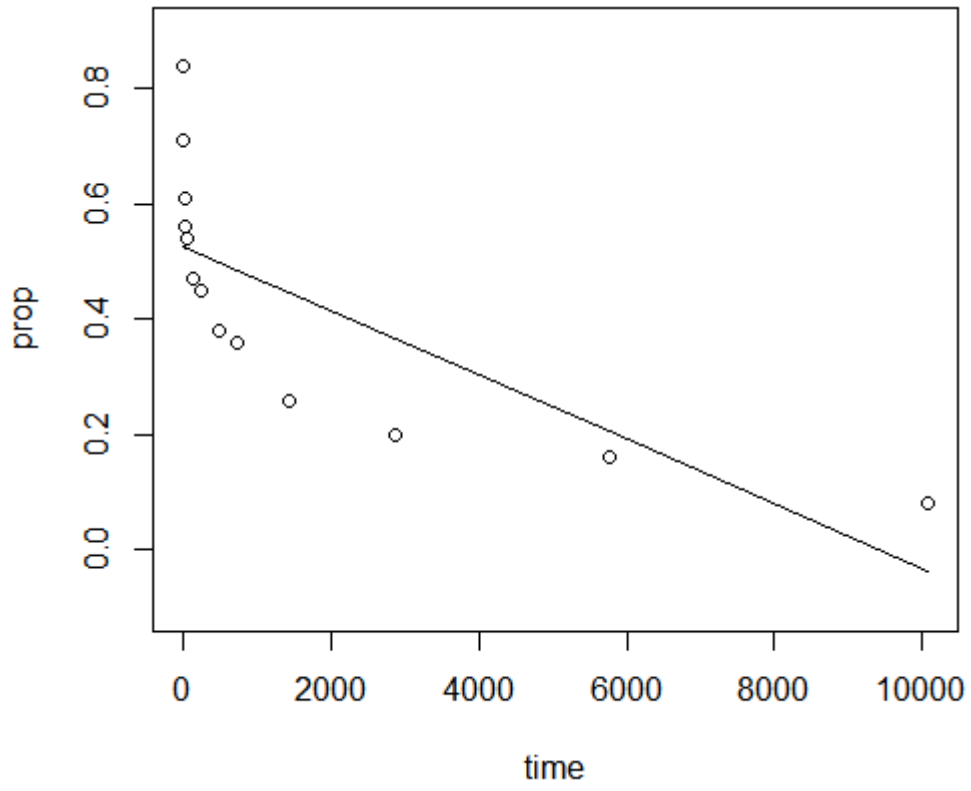
5. Model Summary

```
> summary(model.1)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.259e-01  4.881e-02  10.774 3.49e-07 ***
time        -5.571e-05  1.457e-05  -3.825  0.00282 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1523 on 11 degrees of freedom
Multiple R-squared:  0.5709,  Adjusted R-squared:  0.5318
F-statistic: 14.63 on 1 and 11 DF,  p-value: 0.002817
```

6. Regression Plot

```
plot(x=time, y=prop, ylim=c(-0.1, 0.9),
     panel.last = lines(sort(time),
                        fitted(model.1)[order(time)]))
```

6. Regression Plot

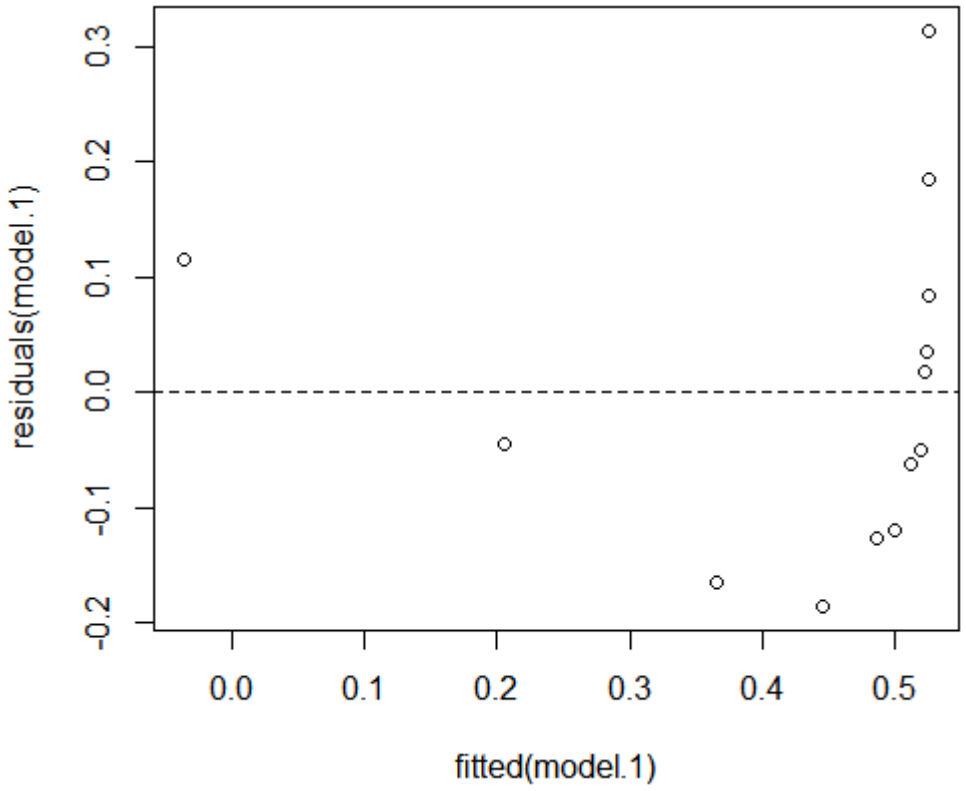


- The resulting fitted line plot suggests that the proportion of recalled items (y) is not linearly related to time (x).

7. Residuals vs. Fits Plot

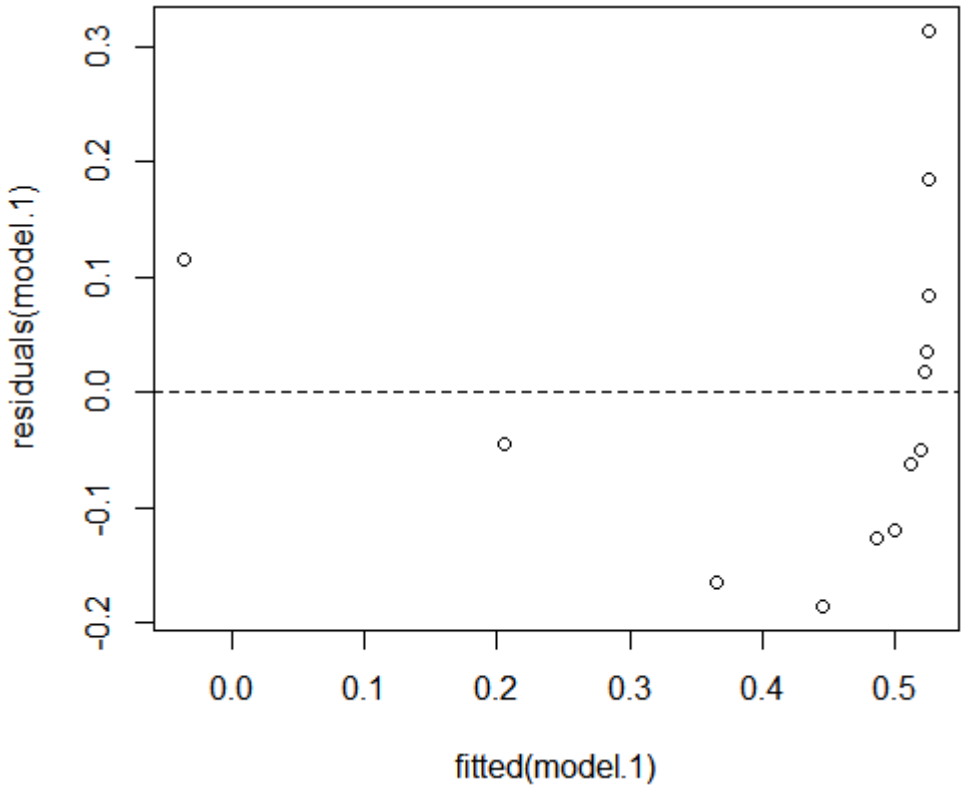
```
plot(x=fitted(model.1), y=residuals(model.1),
     panel.last = abline(h=0, lty=2))
```

7. Residuals vs. Fits Plot



- The residuals vs. fits plot also suggests that the relationship is not linear.

7. Residuals vs. Fits Plot

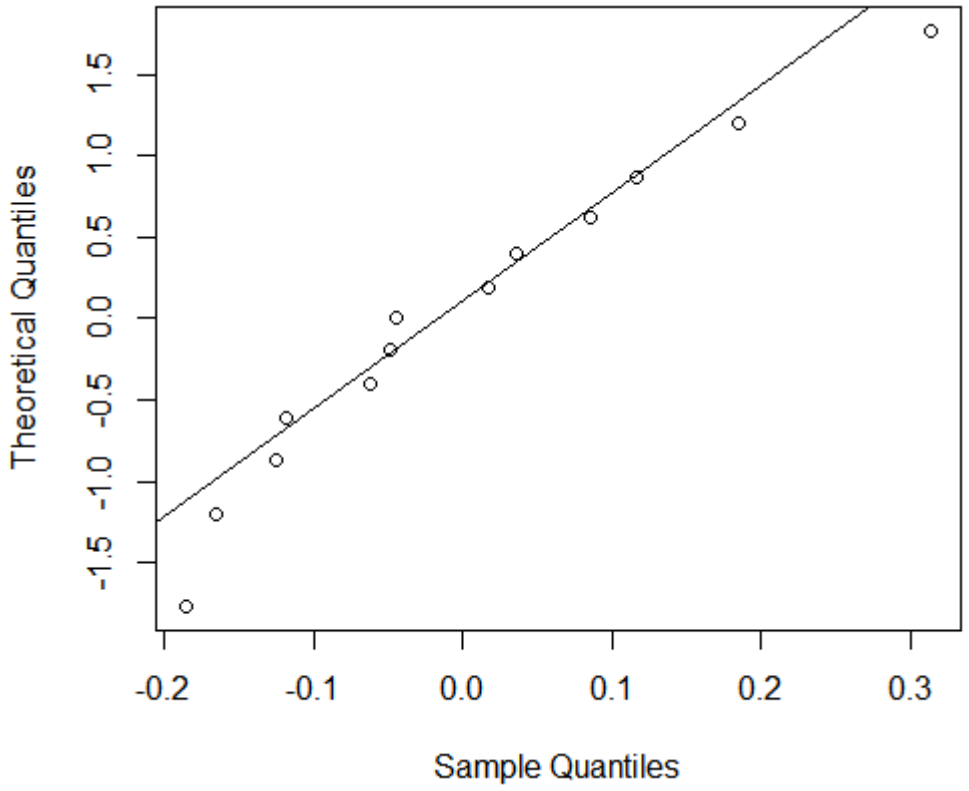


- Because the lack of linearity dominates the plot, we cannot use the plot to evaluate whether or not the error variances are equal. We have to fix the non-linearity problem before we can assess the assumption of equal variances.

8. Normal Probability Plot

```
qqnorm(residuals(model.1), main="", datax=TRUE)
qqline(residuals(model.1), datax=TRUE)
```

8. Normal Probability Plot



- It seems that error terms are normally distributed.

9. Normality Test

```
> shapiro.test(residuals(model.1))

        Shapiro-Wilk normality test

data:  residuals(model.1)
W = 0.94755, p-value = 0.5616
```

10. Evaluation

- In summary, we have a data set in which non-linearity is the only major problem. This situation requests for transforming only the predictor's values.

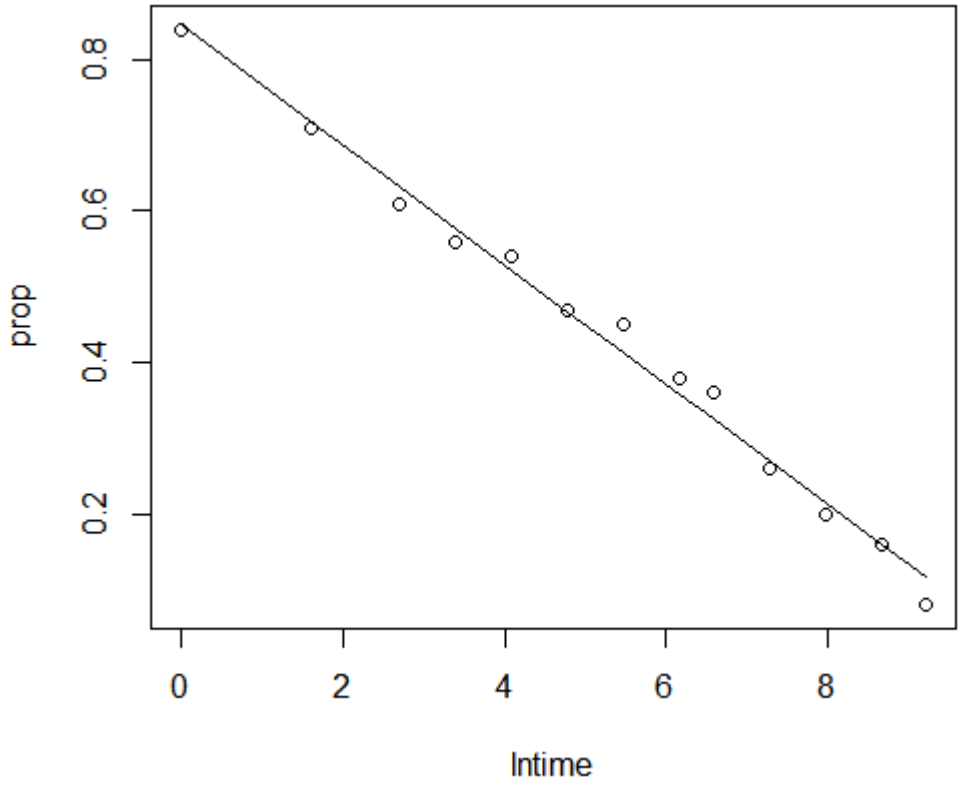
11. Model Creation

```
lntime <- log(time)
model.2 <- lm(prop ~ lntime)
summary(model.2)
```

12. Model Summary

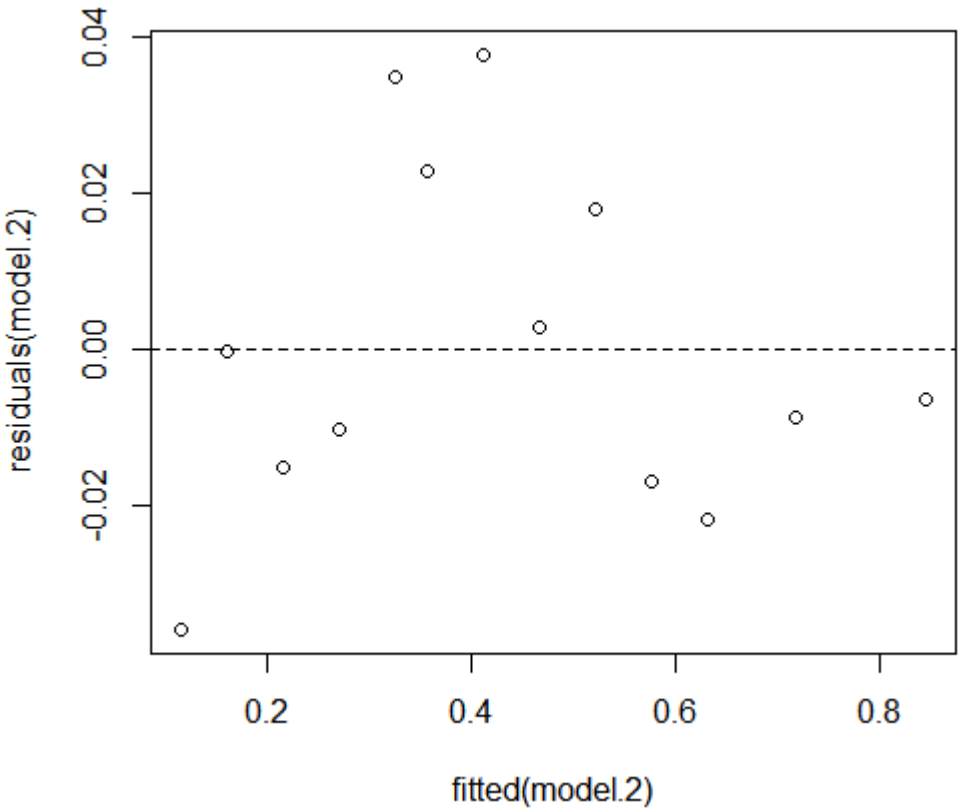
```
> summary(model.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
lntime      -0.079227   0.002416  -32.80 2.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02339 on 11 degrees of freedom
Multiple R-squared:  0.9899,  Adjusted R-squared:  0.989
F-statistic:  1076 on 1 and 11 DF,  p-value: 2.525e-12
```

13. Regression Plot



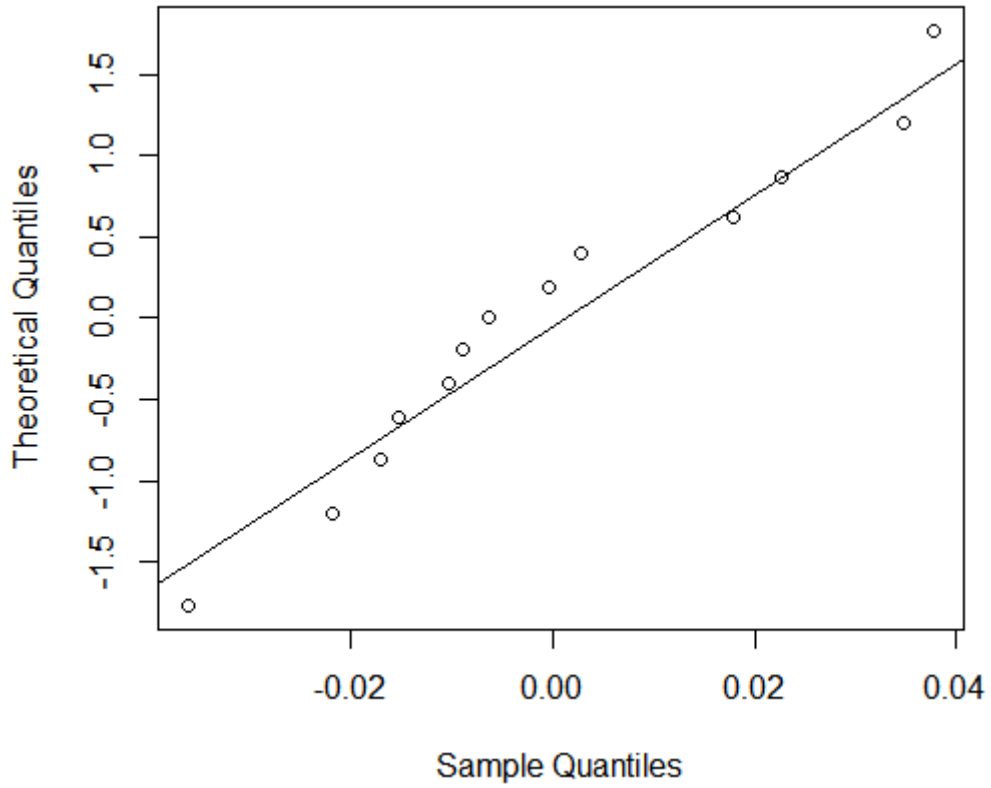
- The resulting fitted line plot suggests that taking the natural logarithm of the predictor values is helpful.

14. Residuals vs. Fits Plot



- The plot also suggests that it is okay to assume that the error variances are equal.

15. Normal Probability Plot



- The normal probability plot of the residuals shows that transforming the x values had no effect on the normality of the error terms.

16. Normality Test

```
> shapiro.test(residuals(model.2))

        Shapiro-Wilk normality test

data:  residuals(model.2)
W = 0.94969, p-value = 0.5936
```

17. What if we had transformed the y values instead?

- The error terms for the memory retention data prior to transforming the x values appear to be well-behaved (in the sense that they appear approximately normal).
- Therefore, we might expect that transforming the y values instead of the x values could cause the error terms to become badly-behaved.

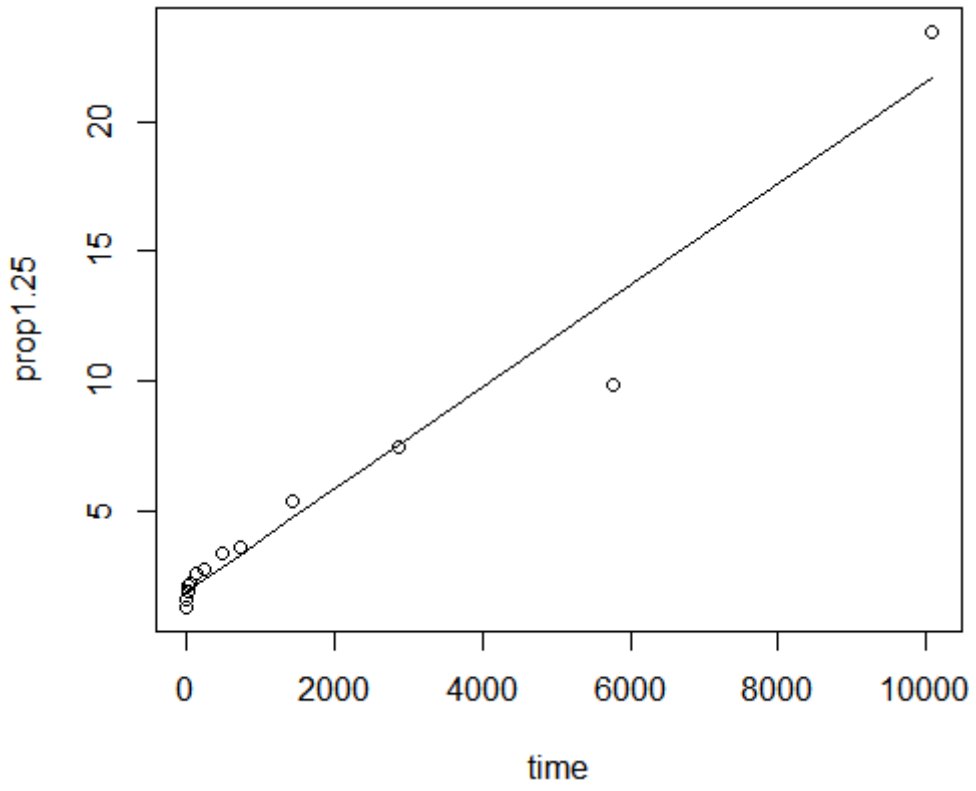
18. Model Creation

```
prop1.25 <- prop^-1.25
model.3 <- lm(prop1.25 ~ time)
summary(model.3)
```

19. Model Summary

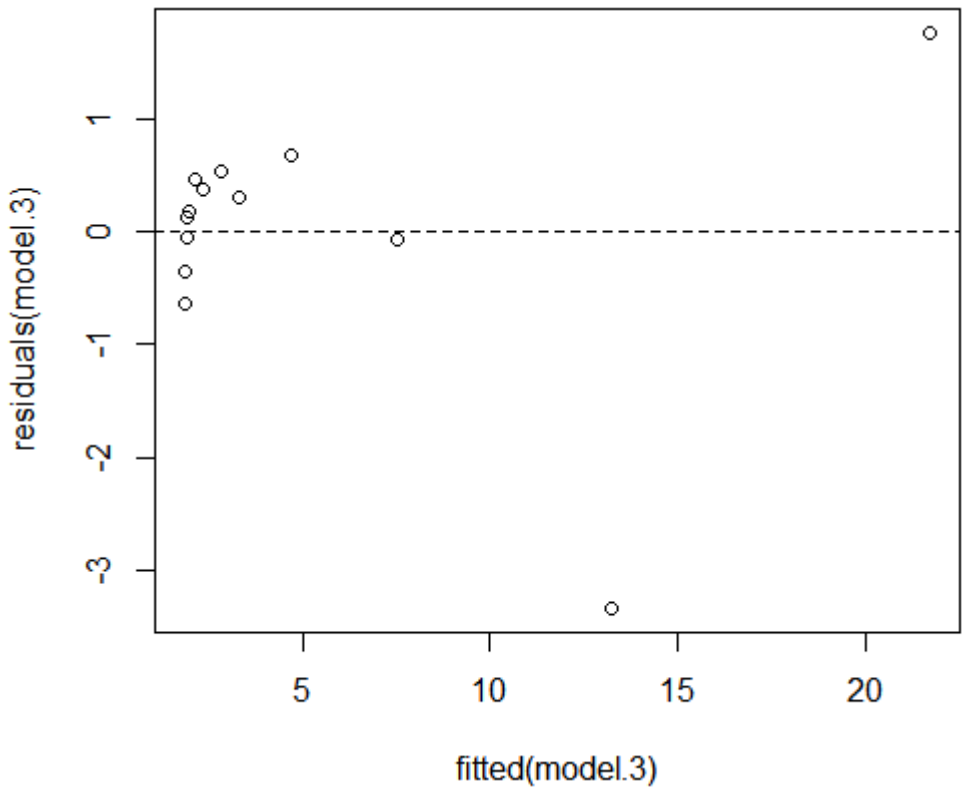
```
> summary(model.3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.8693698  0.3869678   4.831 0.000527 ***
time        0.0019708  0.0001155  17.067 2.91e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.207 on 11 degrees of freedom
Multiple R-squared:  0.9636,  Adjusted R-squared:  0.9603
F-statistic: 291.3 on 1 and 11 DF,  p-value: 2.909e-09
```

20. Regression Plot



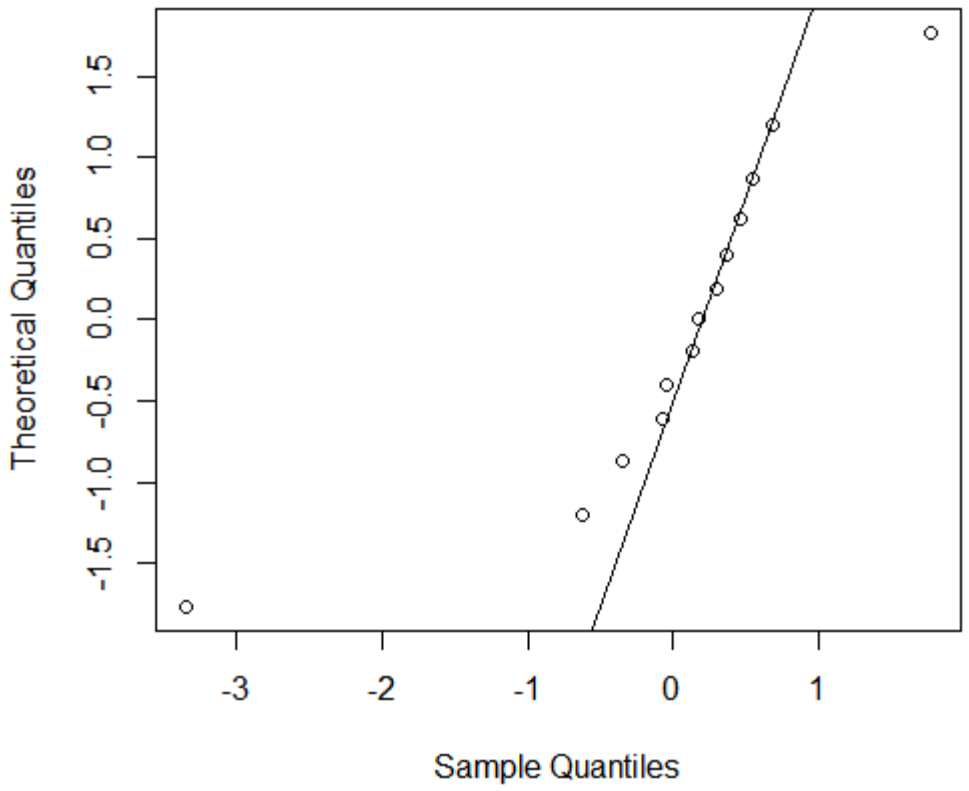
- The fitted line plot illustrates that the transformation does indeed straighten out the relationship — although admittedly not as well as the log transformation of the x values.

21. Residuals vs. Fits Plot



- The residuals show an improvement with respect to non-linearity, although the improvement is not great.

22. Normal Probability Plot



- But now we have non-normal error terms.

23. Normality Test

```
> shapiro.test(residuals(model.3))

        Shapiro-Wilk normality test

data:  residuals(model.3)
W = 0.77665, p-value = 0.003647
```

24. Summary

- If the error terms are well-behaved prior to transformation, transforming the y values can change them into badly-behaved error terms.

25. Question 1

- What is the nature of the association between time since memorized and the effectiveness of recall?
  - The proportion of correctly recalled words is negatively linearly related to the natural log of the time since the words were memorized.

26. Question 2

- Is there an association between time since memorized and effectiveness of recall?

```
> summary(model.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
lntime      -0.079227   0.002416  -32.80 2.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

26. Question 2

- Is there an association between time since memorized and effectiveness of recall?
  - There is significant evidence at the 0.05 level to conclude that there is a linear association between the proportion of words recalled and the natural log of the time since memorized.

27. Question 3

- What proportion of words can we expect a randomly selected person to recall after 1000 minutes?

```
> predict(model.2, interval="prediction",
+         newdata=data.frame(lntime=log(1000)))
        fit       lwr       upr
1 0.2991353 0.2449729 0.3532978
```

28. Question 4

- How much does the expected recall change if time increases ten-fold?
  - $\beta_1 \times \ln 10 = -0.079227 \times \ln 10 = -0.182$

```
> coefficients(model.2)[2]*log(10)
    lntime
-0.1824267
```

28. Question 4

- How much does the expected recall change if time increases ten-fold?
  - We expect the percentage of recalled words to decrease (since the sign is negative) 18.2% for each ten-fold increase in the time since memorization took place.
  - This point estimate is of limited usefulness.

28. Question 4

```
> confint(model.2)
                 2.5 %      97.5 %
(Intercept)  0.81517274  0.87765809
lntime      -0.08454362 -0.07391019

> confint(model.2)[2,]
      2.5 %       97.5 %
-0.08454362 -0.07391019

> confint(model.2)[2,]*log(10)
     2.5 %      97.5 %
-0.1946689 -0.1701845
```

28. Question 4

- We can be 95% confident that the percentage of recalled words will decrease between 17.0% and 19.5% for each ten-fold increase in the time since memorization took place.

2. Transformaing Response

1. Transforming Response

- Transforming the y values should be considered when non-normality and/or unequal variances are the problems with the model. As an added bonus, the transformation on y may also help to "straighten out" a curved relationship.

2. Data: Mammal Gestation

- Data: Mammal Gestation
  - Data on the typical birthweight and length of gestation for various mammals.
  - y (Gestation): length of gestation (in number of days until birth)
  - x (Birthwgt): birthweight (in kg)

3. Data Load

```
> mammgest <- read.table("mammgest.txt", header=T)
> attach(mammgest)
> head(mammgest)
  Row    Mammal Birthwgt Gestation
1   1      Goat     2.75       155
2   2     Sheep     4.00       175
3   3      Deer     0.48       190
4   4 Porcupine     1.50       210
5   5      Bear     0.37       213
6   6     Hippo    50.00       243
```

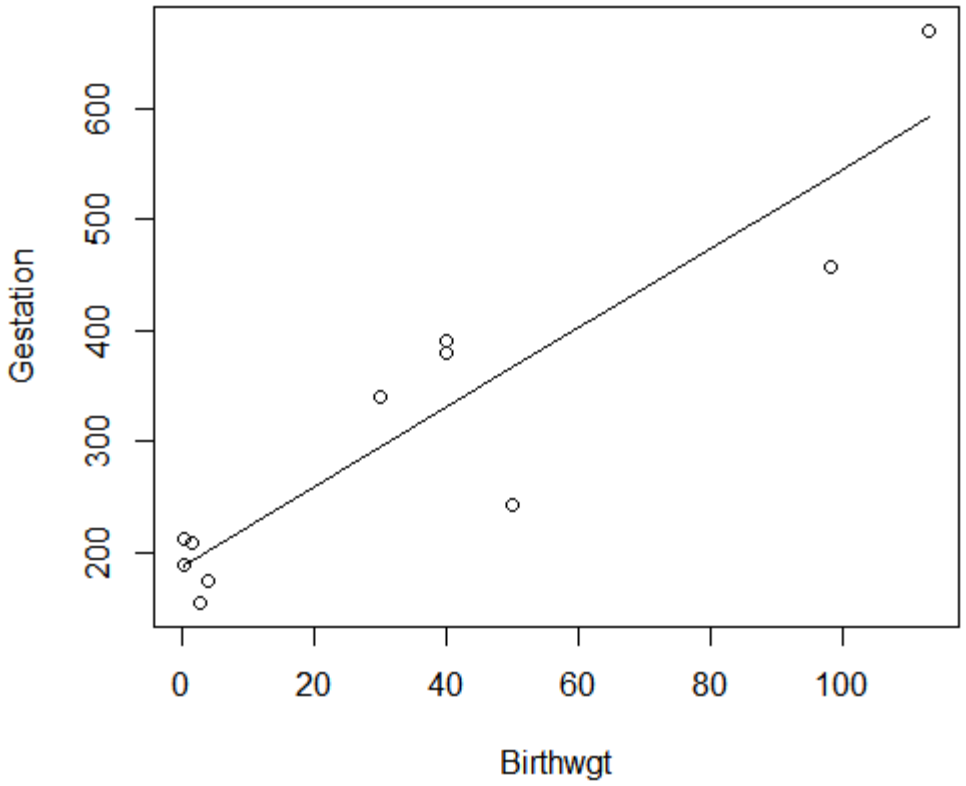
4. Model Creation

```
> model.1 <- lm(Gestation ~ Birthwgt)
> summary(model.1)
```

5. Model Summary

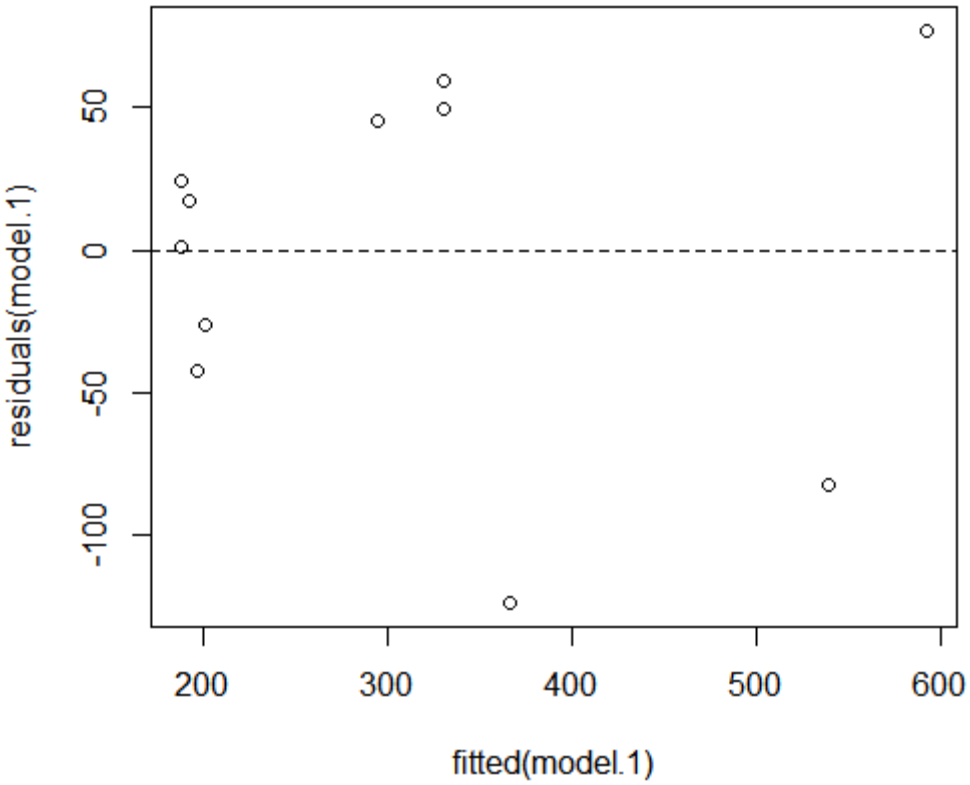
```
> summary(model.1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 187.0837    26.9426   6.944 6.73e-05 ***
Birthwgt      3.5914     0.5247   6.844 7.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.09 on 9 degrees of freedom
Multiple R-squared:  0.8388,  Adjusted R-squared:  0.8209
F-statistic: 46.84 on 1 and 9 DF,  p-value: 7.523e-05
```

6. Regression Plot



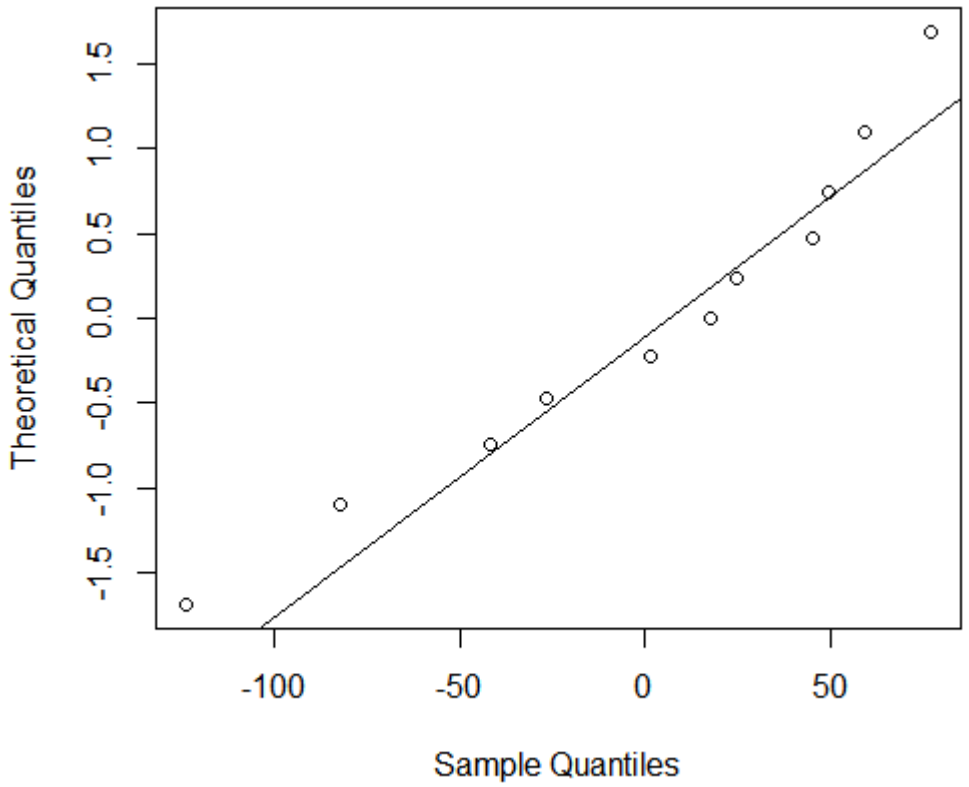
- The fitted line plot suggests that the relationship between gestation length (y) and birthweight (x) is linear.

7. Residuals vs. Fits Plot



- But that the variance of the error terms might not be equal.

8. Normal Probability Plot



- The normal probability plot supports the assumption of normally distributed error terms.

9. Normality Test

```
> shapiro.test(residuals(model.1))

        Shapiro-Wilk normality test

data:  residuals(model.1)
W = 0.93543, p-value = 0.4684
```

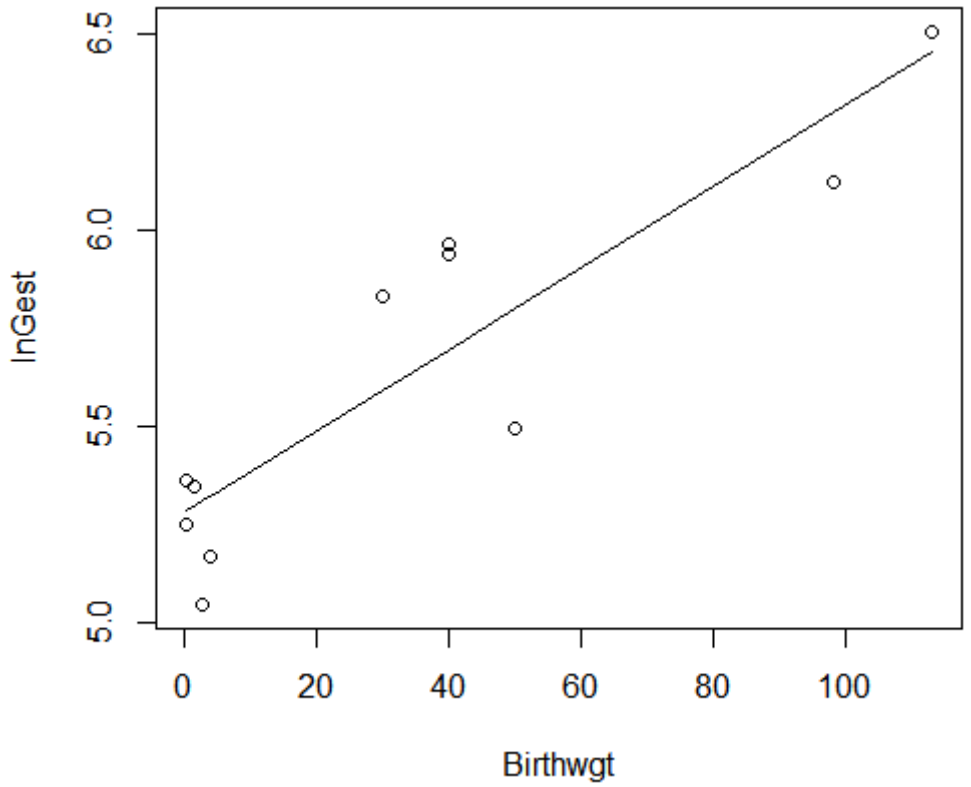
10. Model Creation

```
lnGest <- log(Gestation)
model.2 <- lm(lnGest ~ Birthwgt)
summary(model.2)
```

11. Model Summary

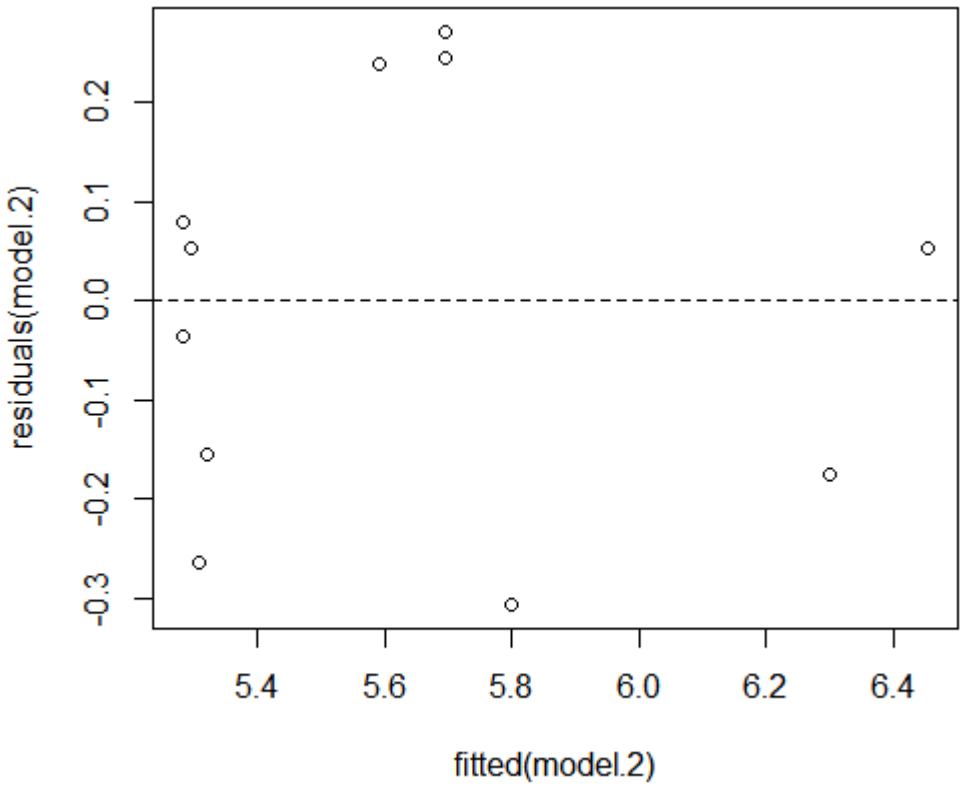
```
> summary(model.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.278817   0.088177  59.866  5.1e-13 ***
Birthwgt    0.010410   0.001717   6.062 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2163 on 9 degrees of freedom
Multiple R-squared:  0.8033,  Adjusted R-squared:  0.7814
F-statistic: 36.75 on 1 and 9 DF,  p-value: 0.0001878
```

12. Regression Plot



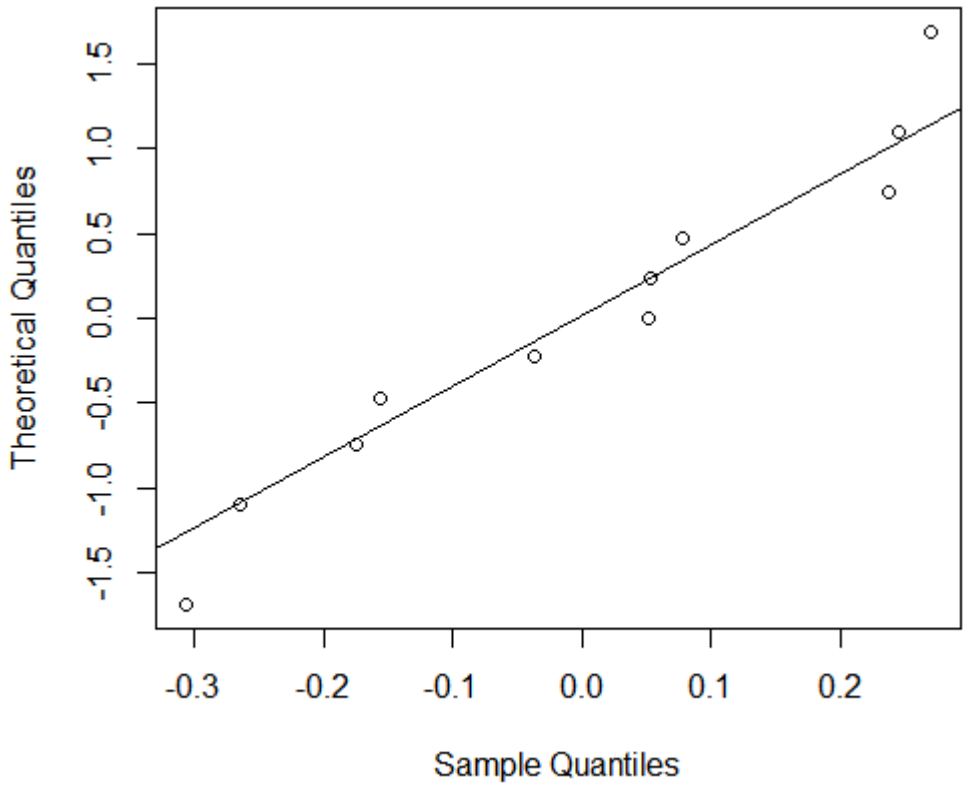
- The resulting fitted line plot suggests that the log transformation of the response does not break linearity.

13. Residuals vs. Fits Plot



- The new residual vs. fits plot shows a marked improvement in the spread of the residuals.

14. Normal Probability Plot



- The log transformation of the response did not adversely affect the normality of the error terms.

15. Normality Test

```
> shapiro.test(residuals(model.2))

        Shapiro-Wilk normality test

data:  residuals(model.2)
W = 0.92739, p-value = 0.385
```

16. Caution

- Note that the r^2 value is lower for the transformed model than for the untransformed model (80.3% versus 83.9%). This does not mean that the untransformed model is preferable. Remember the untransformed model failed to satisfy the equal variance condition, so we should not use this model anyway.

17. Summary

- Transforming the y values should be considered when non-normality and/or unequal variances are the main problems with the model.

18. Question 1

- What is the nature of the association between mammalian birth weight and length of gestation?
  - The natural logarithm of the length of gestation is positively linearly related to birthweight.

19. Question 2

- Is there an association between mammalian birth weight and length of gestation?

```
> summary(model.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.278817   0.088177  59.866  5.1e-13 ***
Birthwgt    0.010410   0.001717   6.062 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

19. Question 2

- Is there an association between time since memorized and effectiveness of recall?
  - There is significant evidence at the 0.05 level to conclude that there is a linear association between the mammalian birthweight and the natural logarithm of the length of gestation.

20. Question 3

- What is the expected gestation length of a new 50 kg mammal?

```
> exp(predict(model.2, interval="prediction",
+             newdata=data.frame(Birthwgt=50)))
       fit      lwr      upr
1 330.0781 197.3013 552.2092
```

21. Question 4

- What is the expected change in length of gestation for each one kg increase in birth weight?
  - $e^{b_1} = e^{0.01041} = 1.01$
  - The predicted gestation changes by a factor of 1.01 for each one unit increase in birthweight.

21. Question 4

```
> exp(coefficients(model.2)[2])
Birthwgt
1.010465

> exp(confint(model.2)[2,])
   2.5 %   97.5 %
1.006547 1.014398
```

21. Question 4

- We can be 95% confident that the length of gestation will increase by a factor between 1.007 and 1.014 for each one kilogram increase in birth weight.

3. Transformaing Predictor and Response

1. Transforming Both the x and y

- This is needed everything seems wrong — when the regression function is not linear and the error terms are not normal and have unequal variances.
  - Transforming the y values corrects problems with the error terms (and may help the non-linearity).
  - Transforming the x values primarily corrects the non-linearity.

2. Data: Short Leaf

- Data: Short Leaf
  - y (Vol): volume of shortleaf pine tree (in cubic feet)
  - x (Diam): diameter of shortleaf pine tree (in inches)

3. Data Load

```
> shortleaf <- read.table("shortleaf.txt", header=T)
> attach(shortleaf)
> head(shortleaf)
  Diam Vol
1  4.4 2.0
2  4.6 2.2
3  5.0 3.0
4  5.1 4.3
5  5.1 3.0
6  5.2 2.9
```

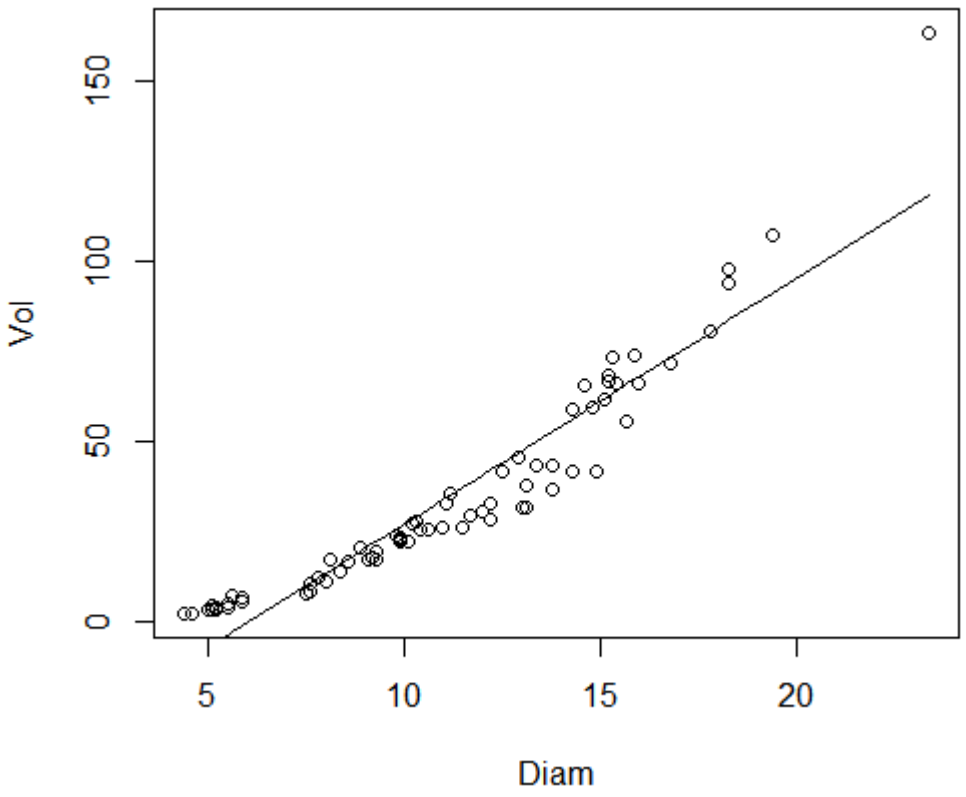
4. Model Creation

```
model.1 <- lm(Vol ~ Diam)
summary(model.1)
```

5. Model Summary

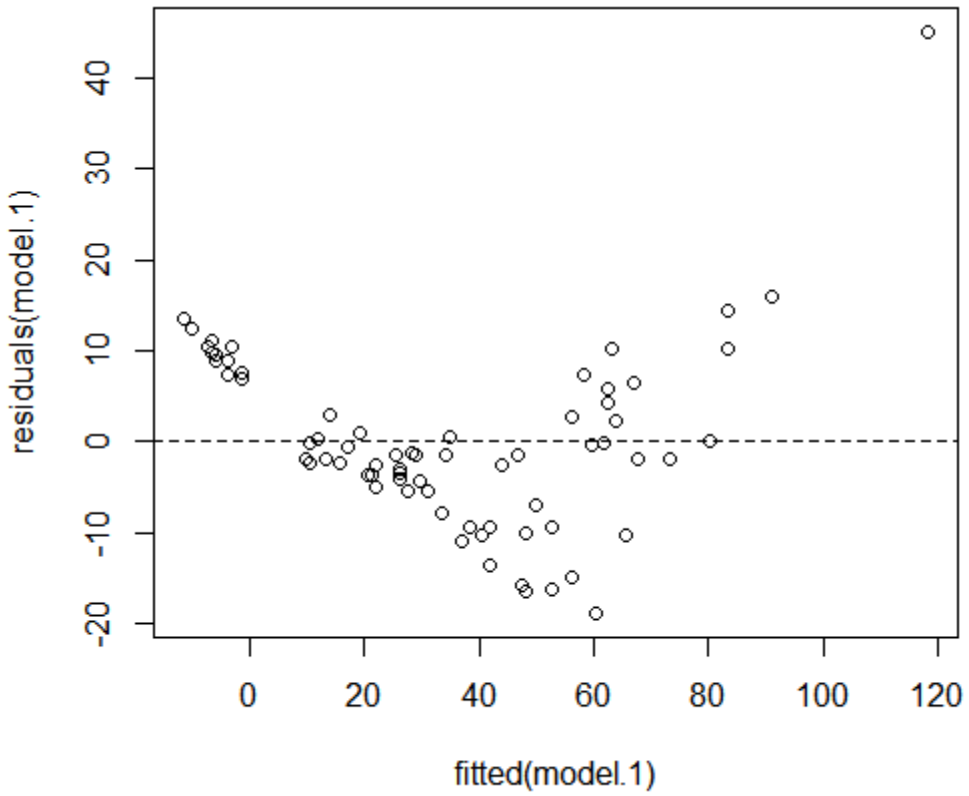
```
> summary(model.1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.5681     3.4269  -12.13   <2e-16 ***
Diam          6.8367     0.2877   23.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.875 on 68 degrees of freedom
Multiple R-squared:  0.8926,  Adjusted R-squared:  0.891
F-statistic: 564.9 on 1 and 68 DF,  p-value: < 2.2e-16
```

6. Regression Plot



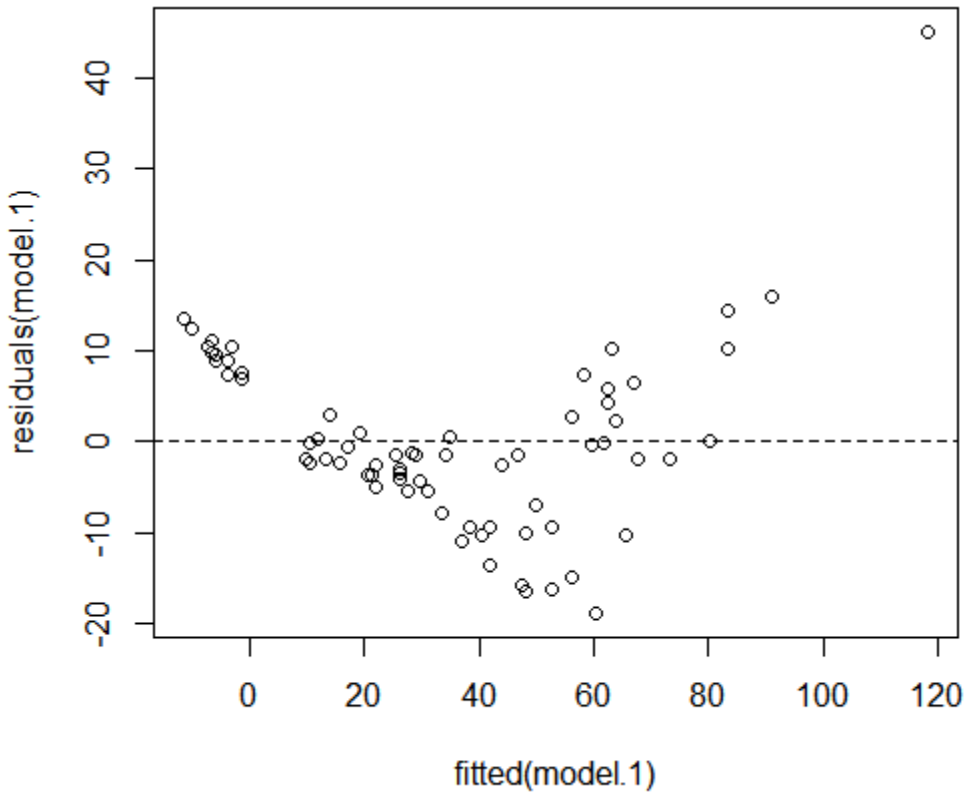
- Although the r^2 value is quite high (89.3%), the fitted line plot suggests that the relationship between tree volume and tree diameter is not linear.

7. Residuals vs. Fits Plot



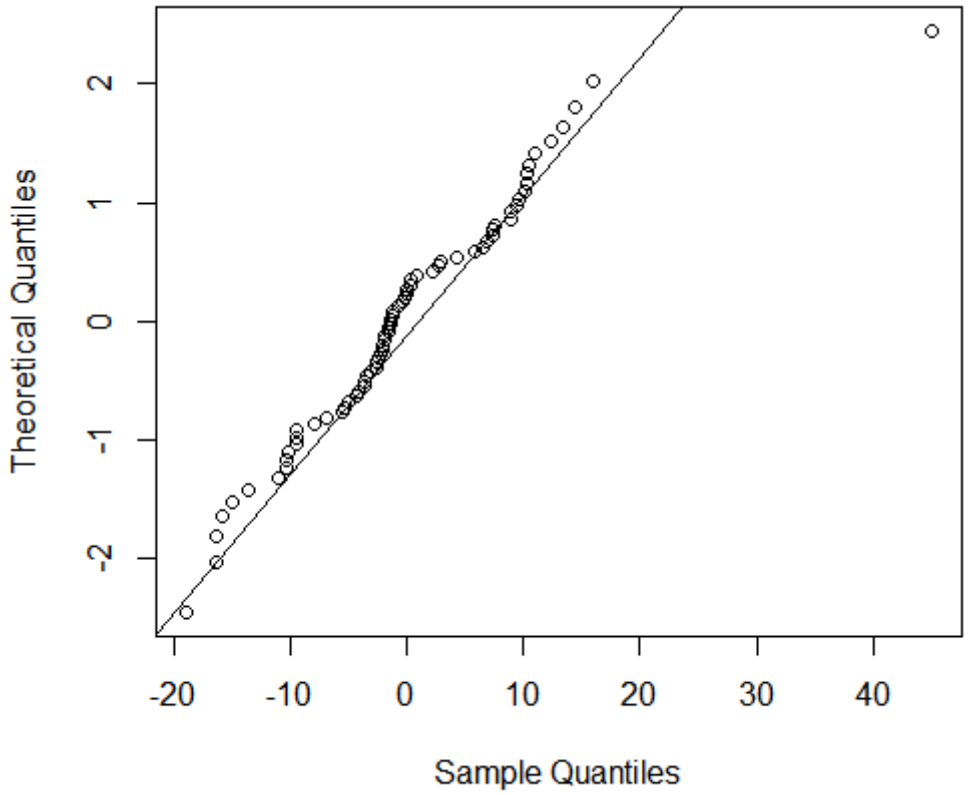
- The residuals vs. fits plot also suggests that the relationship is not linear.

7. Residuals vs. Fits Plot



- Because the lack of linearity dominates the plot, we can not use the plot to evaluate whether the error variances are equal. We have to fix the non-linearity problem before we can assess the assumption of equal variances.

8. Normal Probability Plot



9. Normality Test

```
> shapiro.test(residuals(model.1))

        Shapiro-Wilk normality test

data:  residuals(model.1)
W = 0.91095, p-value = 0.000108
```

10. Evaluation

- It appears as if the relationship between tree diameter and volume is not linear.
- Furthermore, it appears as if the error terms are not normally distributed.

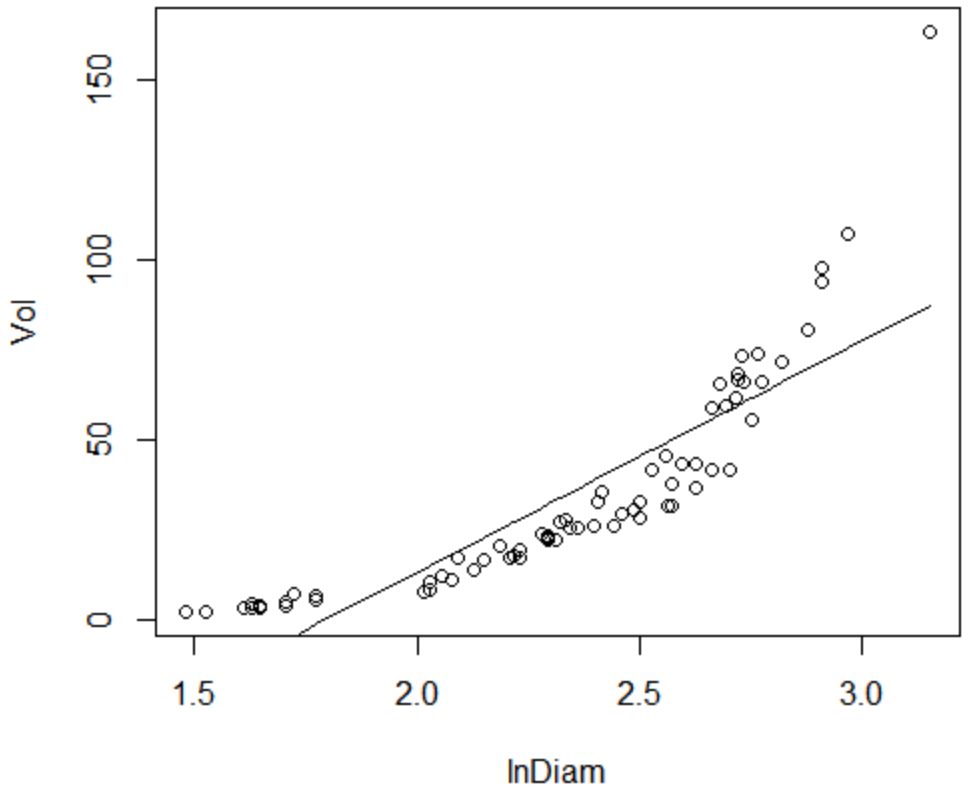
11. Model Creation

```
lnDiam <- log(Diam)
model.2 <- lm(Vol ~ lnDiam)
summary(model.2)
```

12. Model Summary

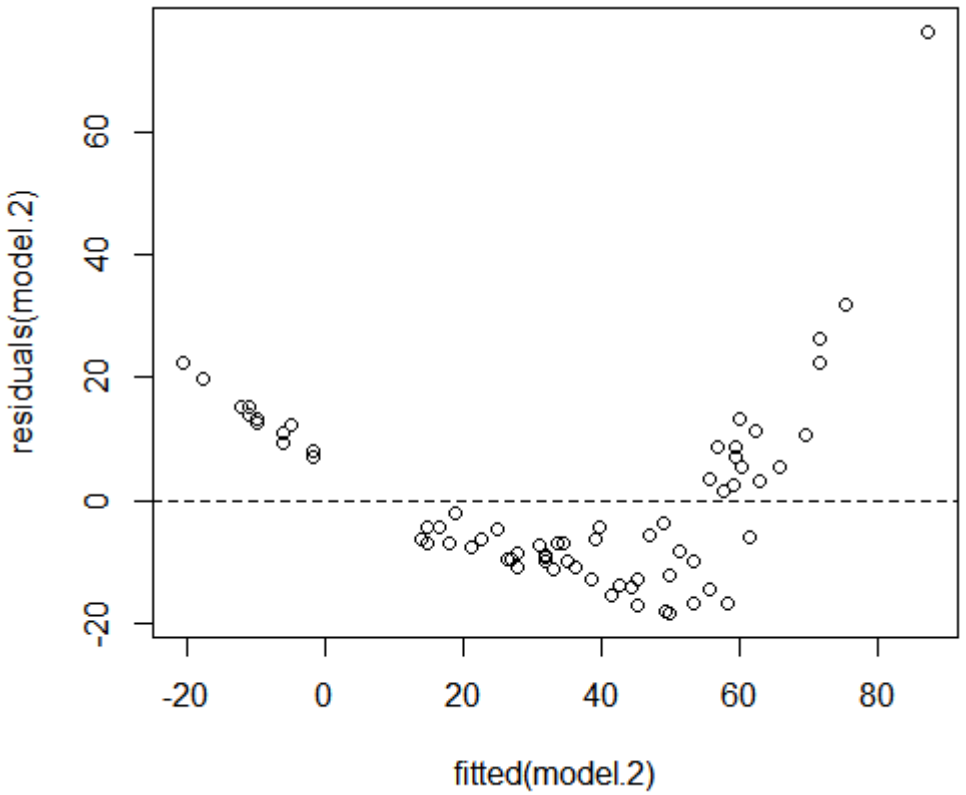
```
> summary(model.2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -116.162     10.830  -10.73 2.88e-16 ***
lnDiam        64.536      4.562   14.15  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.17 on 68 degrees of freedom
Multiple R-squared:  0.7464,  Adjusted R-squared:  0.7427
F-statistic: 200.2 on 1 and 68 DF,  p-value: < 2.2e-16
```

13. Regression Plot



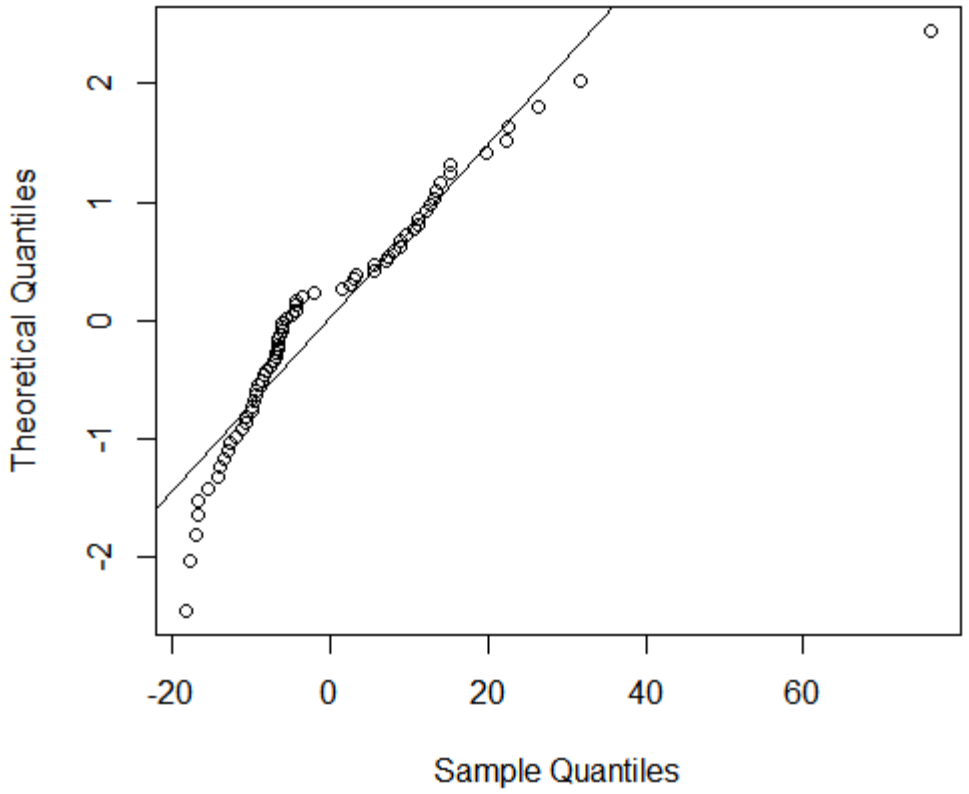
- The fitted line plot suggests that the relationship is still not linear.

14. Residuals vs. Fits Plot



- The residuals vs. fits plot also still suggests a non-linear relationship.

15. Normal Probability Plot



- There is little improvement in the normality of the error terms.

16. Normality Test

```
> shapiro.test(residuals(model.2))

        Shapiro-Wilk normality test

data:  residuals(model.2)
W = 0.8269, p-value = 1.336e-07
```

17. Evaluation

- So, transforming x alone didn't help much. Let's also try transforming the response (y) values. In particular, let's take the natural logarithm of the tree volumes.

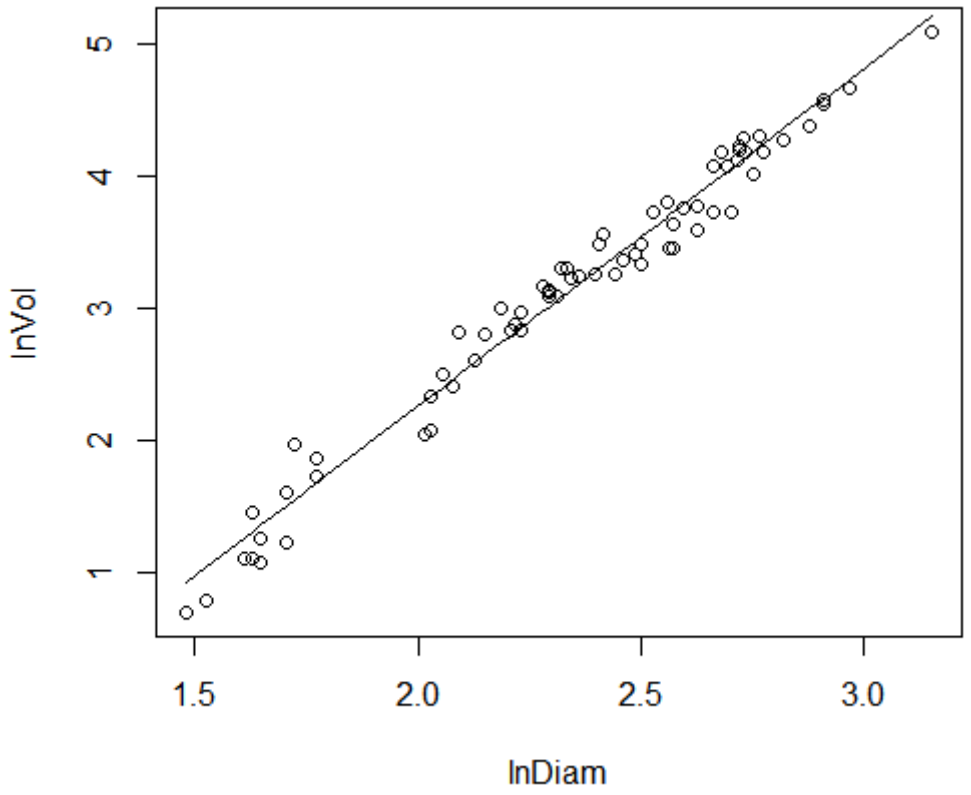
18. Model Creation

```
lnVol <- log(Vol)
model.3 <- lm(lnVol ~ lnDiam)
summary(model.3)
```

19. Model Summary

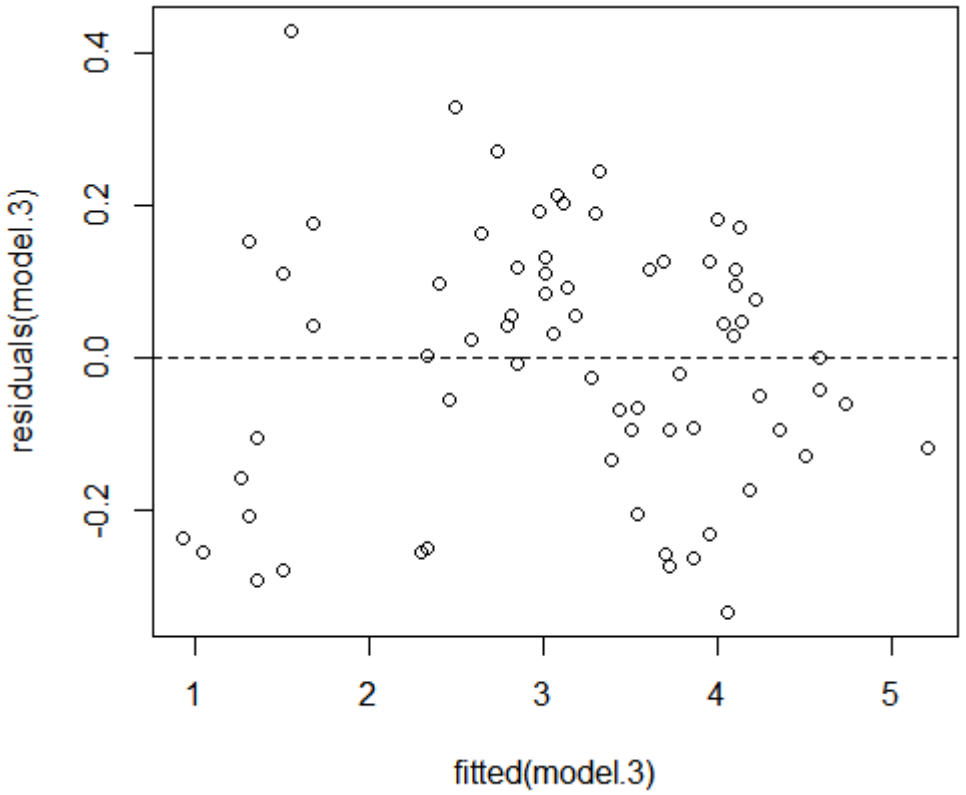
```
> summary(model.3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.8718     0.1216  -23.63   <2e-16 ***
lnDiam        2.5644     0.0512   50.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1703 on 68 degrees of freedom
Multiple R-squared:  0.9736,  Adjusted R-squared:  0.9732
F-statistic:  2509 on 1 and 68 DF,  p-value: < 2.2e-16
```

20. Regression Plot



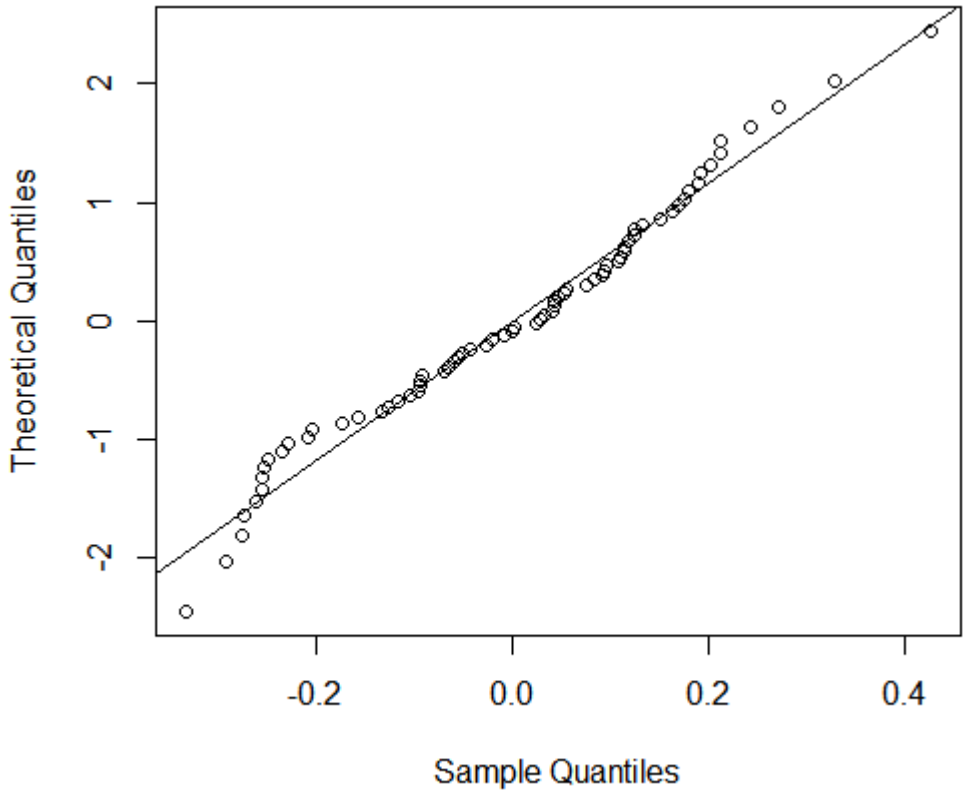
- The relationship between the natural log of the diameter and the natural log of the volume looks linear and strong.

21. Residuals vs. Fits Plot



- The residuals vs. fits plot provides yet more evidence of a linear relationship between lnVol and lnDiam.
- The residuals bounce randomly around the residual = 0 line.

22. Normal Probability Plot



- The normal probability plot has improved substantially.

23. Normality Test

```
> shapiro.test(residuals(model.3))

        Shapiro-Wilk normality test

data:  residuals(model.3)
W = 0.9767, p-value = 0.2164
```

24. Summary

- In summary, it appears as if the model with the natural log of tree volume as the response and the natural log of tree diameter as the predictor works well. The relationship appears to be linear and the error terms appear independent and normally distributed with equal variances.

25. Question 1

- What is the nature of the association between diameter and volume of shortleaf pines?
  - The natural logarithm of tree volume is positively linearly related to the natural logarithm of tree diameter.

26. Question 2

- Is there an association between the diameter and volume of shortleaf pines?

```
> summary(model.3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.8718     0.1216  -23.63   <2e-16 ***
lnDiam        2.5644     0.0512   50.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

26. Question 2

- Is there an association between the diameter and volume of shortleaf pines?
  - There is significant evidence at the 0.01 level to conclude that there is a linear association between the natural logarithm of tree volume and the natural logarithm of tree diameter.

27. Question 3

- What is the "average" volume of all shortleaf pine trees that are 10" in diameter?

```
> exp(predict(model.3, interval="confidence",
+             newdata=data.frame(lnDiam=log(10))))
       fit       lwr       upr
1 20.75934 19.92952 21.62372
```

28. Question 4

- What is the expected change in volume for a two-fold increase in diameter?
  - $\ln \Delta y = b_1(\ln 2x - \ln x) = b_1 \ln 2 = \ln 2^{b_1}$
  - $2^{b_1} = \Delta y$

28. Question 4

```
> 2^(coefficients(model.3)[2])
  lnDiam
5.915155

> 2^(confint(model.3)[2,])
   2.5 %   97.5 %
5.510776 6.349207
```

28. Question 4

- We can be 95% confident that the volume will increase by a factor between 5.51 and 6.35 for each two-fold increase in diameter.

4. Other Data Transformations

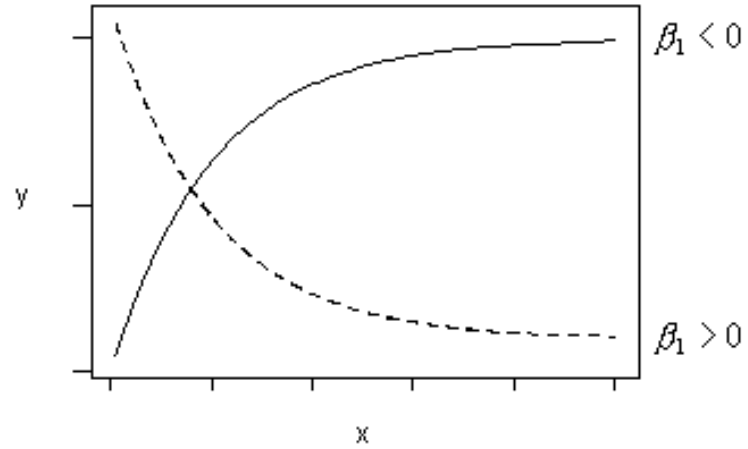
1. Other Transformations

- There are other transformations you could try in an attempt to correct problems with your model.
- One thing to keep in mind though is that transforming your data almost always involves lots of trial and error.

2. Residual Plot and Transformations

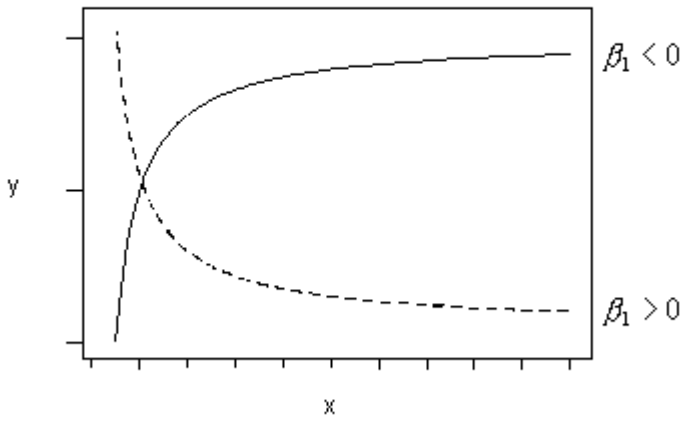
- If the primary problem with your model is non-linearity, look at a scatter plot of the data to suggest transformations that might help. (This only works for simple linear regression models with a single predictor. For multiple linear regression models, look at residual plots instead.)

3. Pattern 1



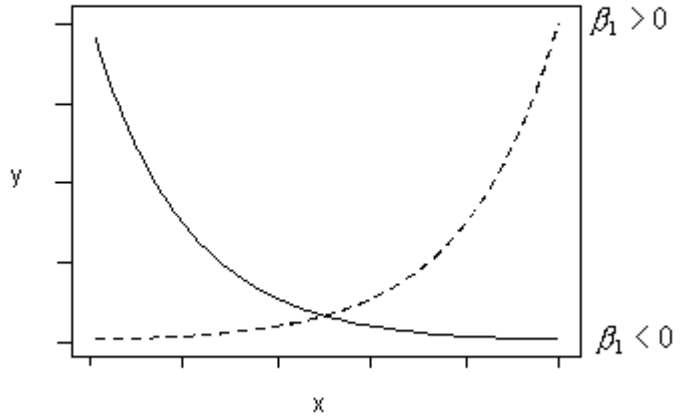
- If the trend in your data follows either of these patterns, you could try fitting this regression function.
  - $\mu_Y = \beta_0 + \beta_1 e^{-x}$

4. Pattern 2



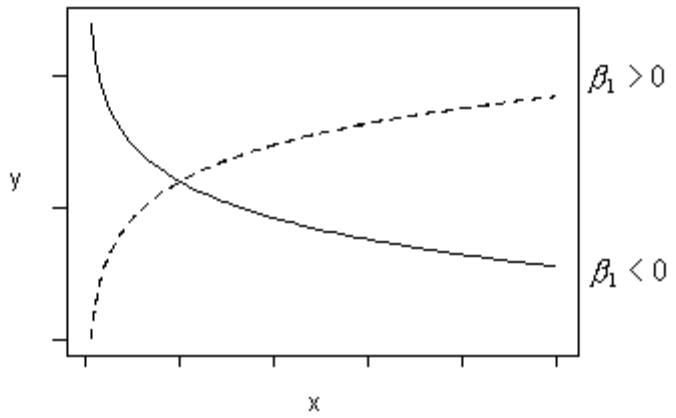
- If the trend in your data follows either of these patterns, you could try fitting this regression function.
  - $\mu_Y = \beta_0 + \beta_1 \frac{1}{x}$
- This is sometimes called a "reciprocal" transformation.

5. Pattern 3



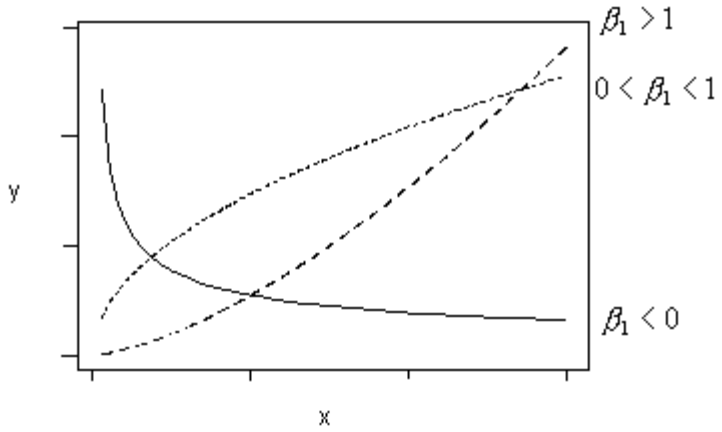
- If the trend in your data follows either of these patterns, you could try fitting this regression function.
  - $\mu_{\ln Y} = \beta_0 + \beta_1 x$

6. Pattern 4



- If the trend in your data follows either of these patterns, you could try fitting this regression function.
  - $\mu_Y = \beta_0 + \beta_1 \ln x$

7. Pattern 5



- If the trend in your data follows either of these patterns, you could try fitting this regression function.
  - $\mu_{\ln Y} = \beta_0 + \beta_1 \ln x$

8. Power Transformation on y

- If the variances are unequal and/or error terms are not normal, try a "power transformation" on y.
- A power transformation on y involves transforming the response by taking it to some power λ. $y^* = y^\lambda$

5. Summary

1. What to Try?

- When there is curvature in the data, there might possibly be some theory in the literature of the subject matter to suggests an appropriate equation.
- Or, you might have to use trial and error data exploration to determine a model that fits the data.

1. What to Try?

- In the trial and error approach, you might try polynomial models or transformations of the x-variable(s) and/or y-variable such as square root, logarithmic, or reciprocal transformations.

2. Transform Predictors or Response?

- If you transform the y-variable, you will change the variance of the y-variable and the errors. You may wish to try transformations of the y-variable (e.g. $\ln y$, $\sqrt{y}$, $\frac{1}{y}$) when there is evidence of nonnormality and/or nonconstant variance problems in one or more residual plots.

2. Transform Predictors or Response?

- Try transformations of the x-variable(s) (e.g. $\ln x$, $\frac{1}{x}$, x^2) when there are strong nonlinear trends in one or more residual plots.

3. Why Might Logarithms Work?

- Logarithms are often used because they are connected to common exponential growth and power curve relationships.

3. Why Might Logarithms Work?

- Exponential growth equation
  - $y = a \times e^{bx}$
  - $\ln y = \ln a + bx$

3. Why Might Logarithms Work?

- Power curve equation
  - $y = a \times x^b$
  - $\ln y = \ln a + b \ln x$

Next

Chapter 12
Model Building