Remedial Procedure in SLR: transformation

Overview of remedial measures

If the simple linear regression model is not appropriate for a data set

- Abandon regression model and develop a more appropriate model
- Employ some transformation on the data so that regression model is appropriate *for the transformed data*

  - Nonlinearity of regression function → Transformations
  - Non-constancy of error variance → Transformations and Weighted least squares
  - Non-normality of error terms → Transformations
  - Outliers → Transformations or Robust regression
  - Non-independence of Error terms → Autocorrelation, time series analysis

When the error terms approximately have a Normal distribution with constant variance

- Transformation on X should be attempted (at first).

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, where\ X_1 = X, X_2 = X^2$$

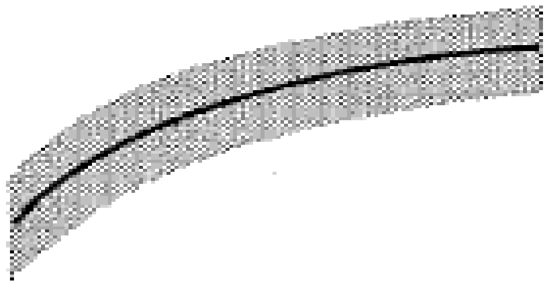
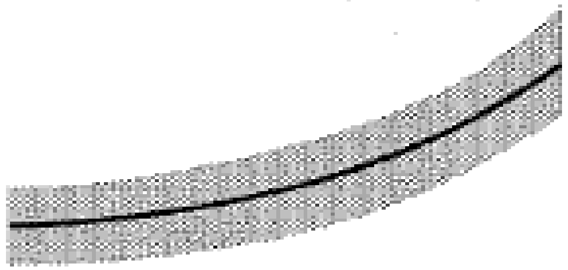
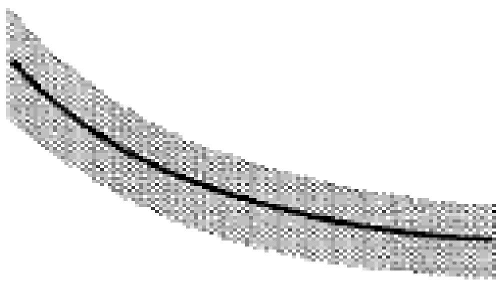
$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \varepsilon, where\ X_1 = \log(X)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, where\ X_3 = X_1 X_2$$

- The reason why transformation on Y may not be desirable is that transformation of Y may materially change the shape of distribution of the error terms from the Normal distribution and lead to differing error term variances.

$$Y = \beta_0 + \beta_1 X + \varepsilon \rightarrow \sqrt{Y} = \beta_0 + \beta_1 X_1 + new\ \varepsilon$$

Some common transformation form on X

	Prototype Regression Pattern	Transformations of X	
(a)		$X' = \log_{10} X$	$X' = \sqrt{X}$
(b)		$X' = X^2$	$X' = \exp(X)$
(c)		$X' = 1/X$	$X' = \exp(-X)$

Comment

- If some of the X data are near 0 and reciprocal transformation is desired. Shift the origin by

$$X' = \frac{1}{X+k}$$

Where $k \neq 0$

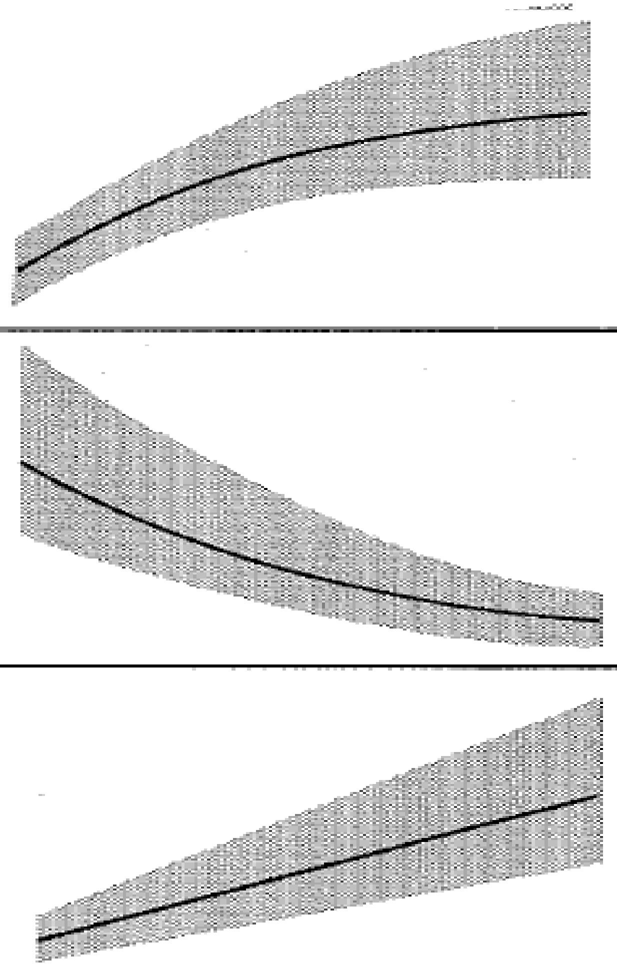
Unequal error variance and nonnormality of the error terms frequently appear together, and we need a transformation on Y

Prototype Regression Pattern

(a)

(b)

(c)



Transformations on Y

$Y' = Y^{\lambda}$ (Box-Cox Transformation)

For example,

$$\lambda = 2 \quad Y' = Y^2$$
$$\lambda = 0.5 \quad Y' = \sqrt{Y}$$
$$\boxed{\lambda = 0 \quad Y' = \ln Y}$$
$$\lambda = -0.5 \quad Y' = \frac{1}{\sqrt{Y}}$$
$$\lambda = -1 \quad Y' = \frac{1}{Y}$$

Comment

- Consider use constant values to validate the transformation function

$$Y' = \log_{10}(Y + k)$$
$$k \text{ is selected such that } Y + k > 0 \text{ for all } Y.$$

- Can be combined with transformation on X

Box-Cox Procedure

Transformations on Y sometimes help with variance issue: non-normality and non-constant.

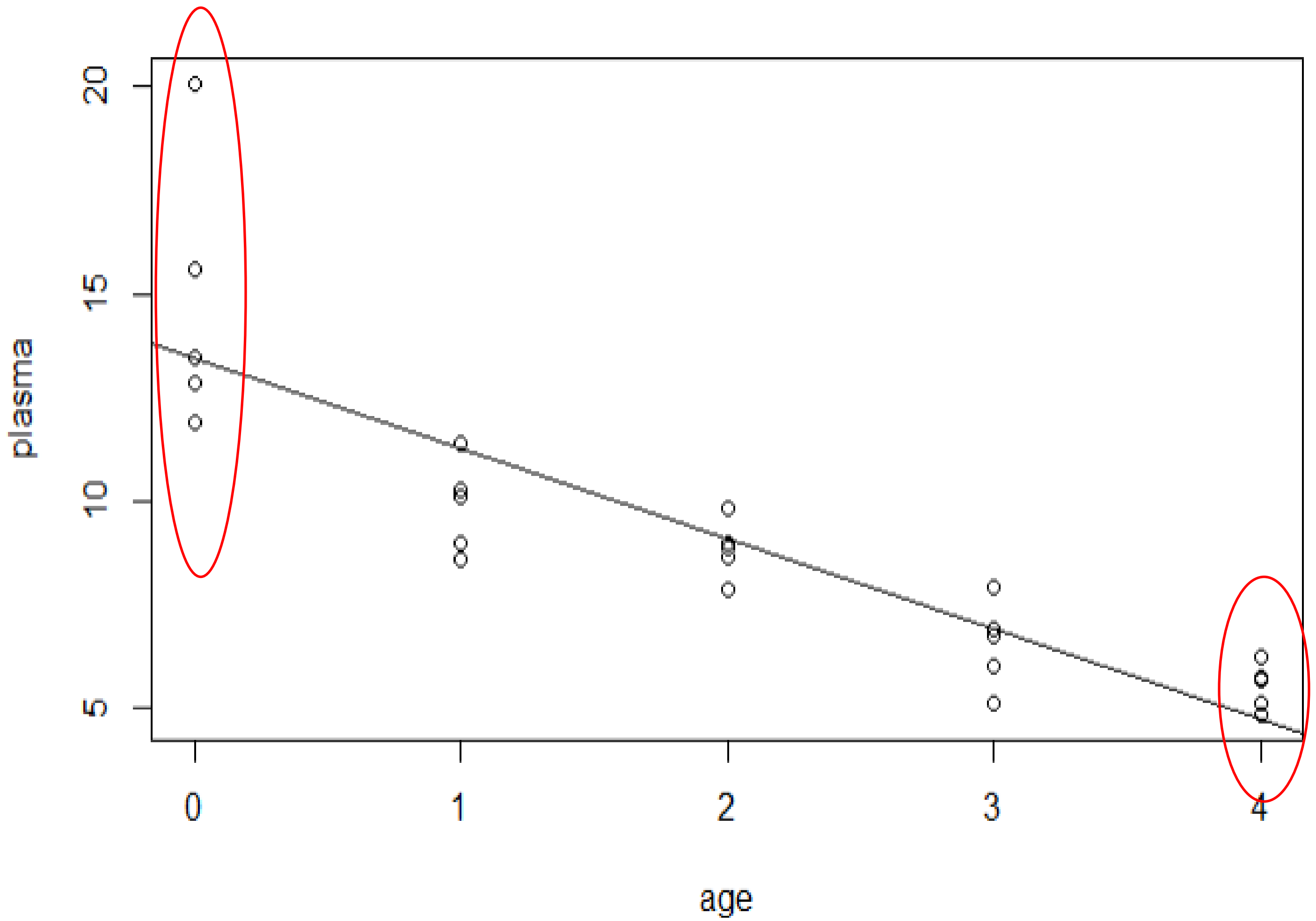
- Box-Cox considers a family of so-called “power transformations”,

$$Y' = Y^{\lambda}$$

- “Works by using the method of maximum likelihood or minimum SSE to find the value of λ that produces the best (transformed) regression $Y^{\lambda} = \beta_0 + \beta X + \varepsilon$
- Need to check assumptions for the transformed regression model.

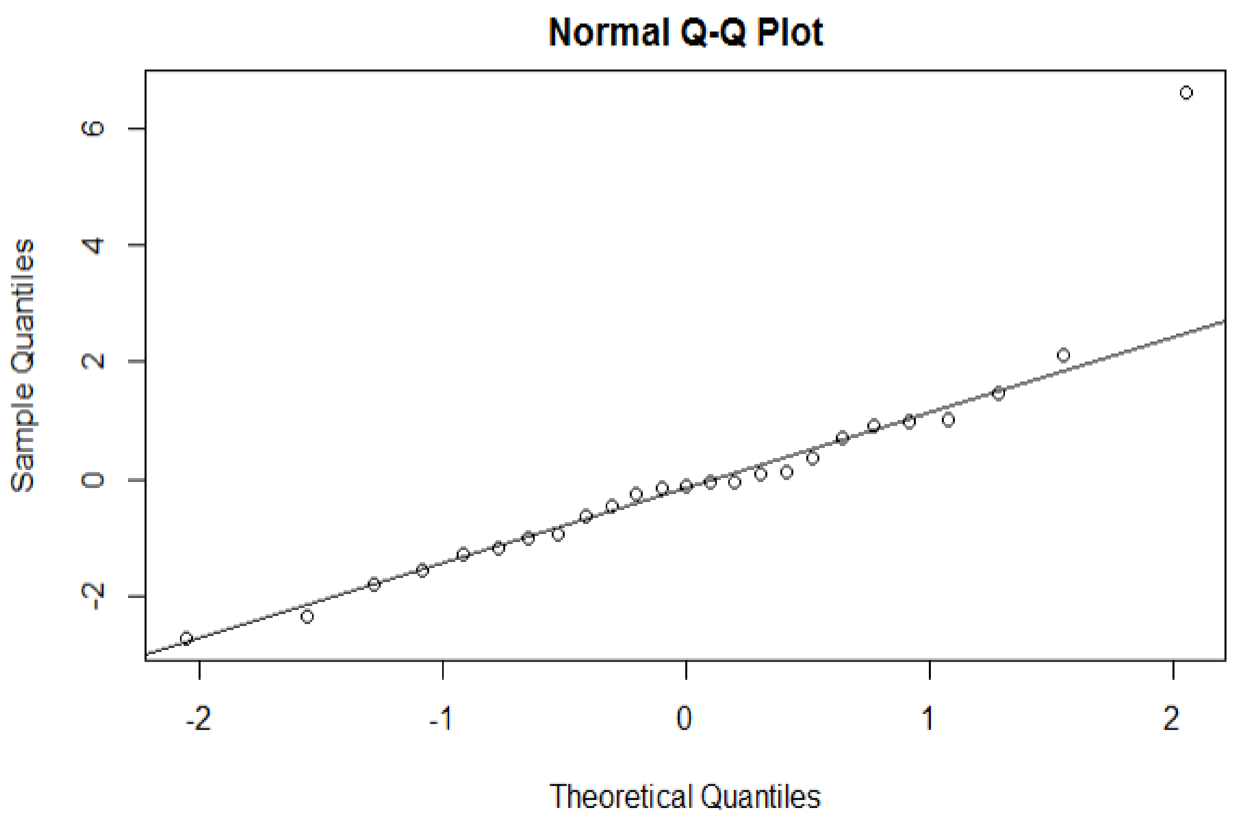
The Plasma example

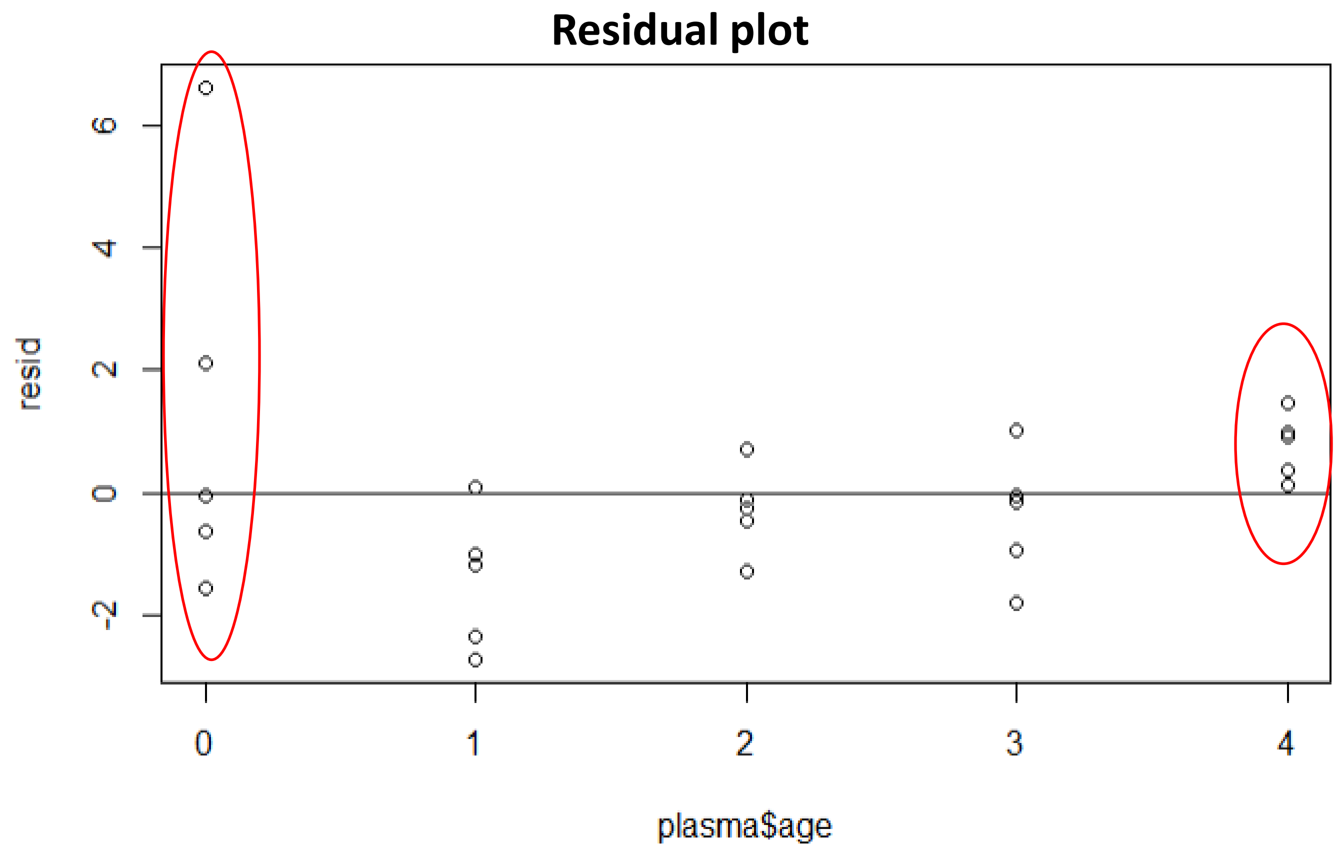
Age (X) and plasma level of a poly amine (Y) for a portion of the 25 healthy children are studied.
Scatter plot shows there is greater variability for younger children than for older ones




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.4752     0.6379  21.126  < 2e-16 ***
age          -2.1820     0.2604  -8.379 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.841 on 23 degrees of freedom
Multiple R-squared:  0.7532,    Adjusted R-squared:  0.7425
F-statistic: 70.21 on 1 and 23 DF,  p-value: 1.92e-08
```

Check Normality and constancy on the residuals





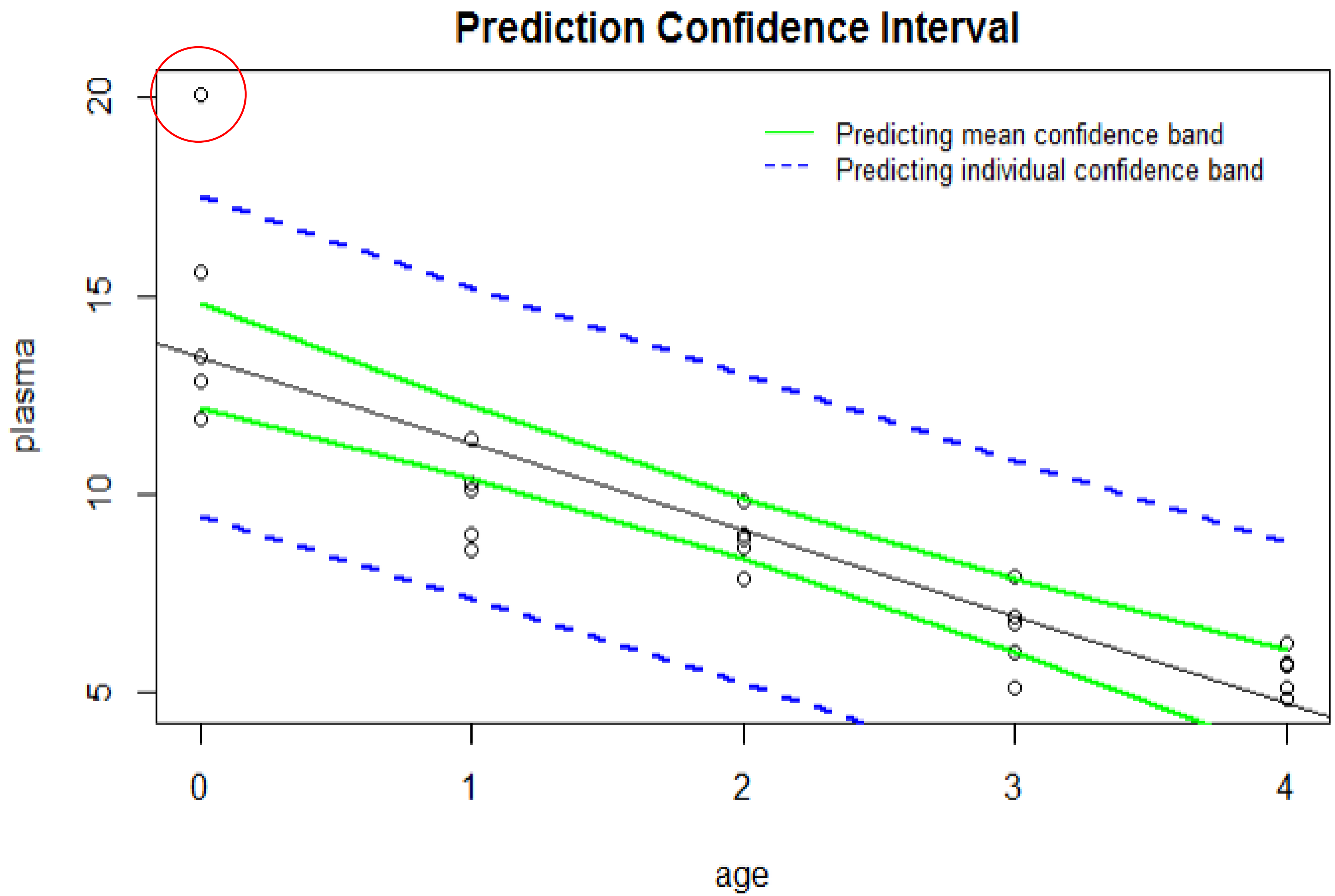

```
        Shapiro-Wilk normality test

data:  residuals(plasma.mod)
W = 0.83903, p-value = 0.001098
```

```
shapiro.test(residuals(plasma.mod))
qqnorm(residuals(plasma.mod))
qqline(residuals(plasma.mod))
```

```
  Brown-Forsythe Test
-------------------------------------------------------
  data : residual and agef

  statistic   : 2.059299
  num df      : 4
  denom df    : 6.526859
  p.value     : 0.1965498

  Result      : Difference is not statistically significant.
-------------------------------------------------------
```

```
bf.test(residual~agef, plasma)
```

Confident interval band



$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}}$

Where $s^2_{\{\hat{Y}_h\}} = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$

$\hat{Y}_h \pm t_c s_{\{pred\}}$

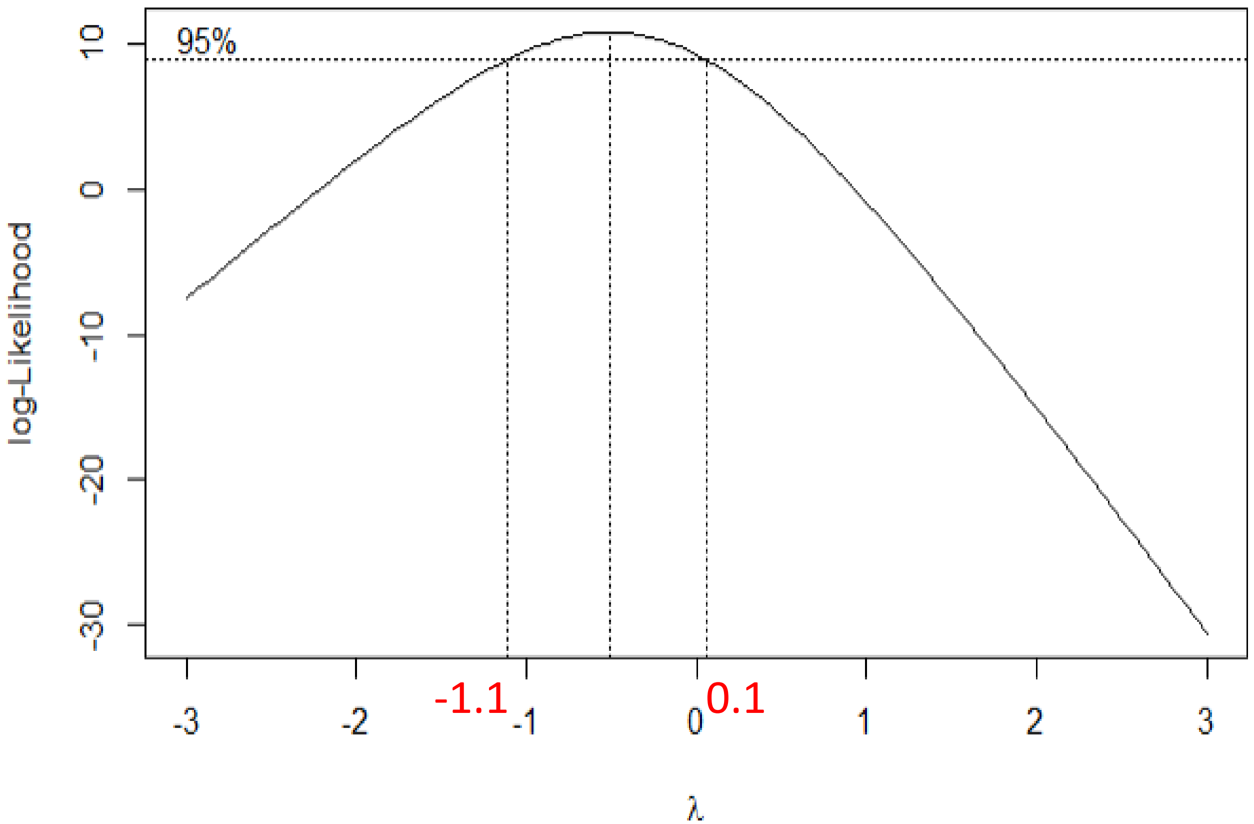
Where $s^2_{\{pred\}} = s^2 + s^2_{\{\hat{Y}_h\}}$

After a careful exam on the experiment procedure, no mistake has been found, hence we should keep this observation.

Box-cox procedure

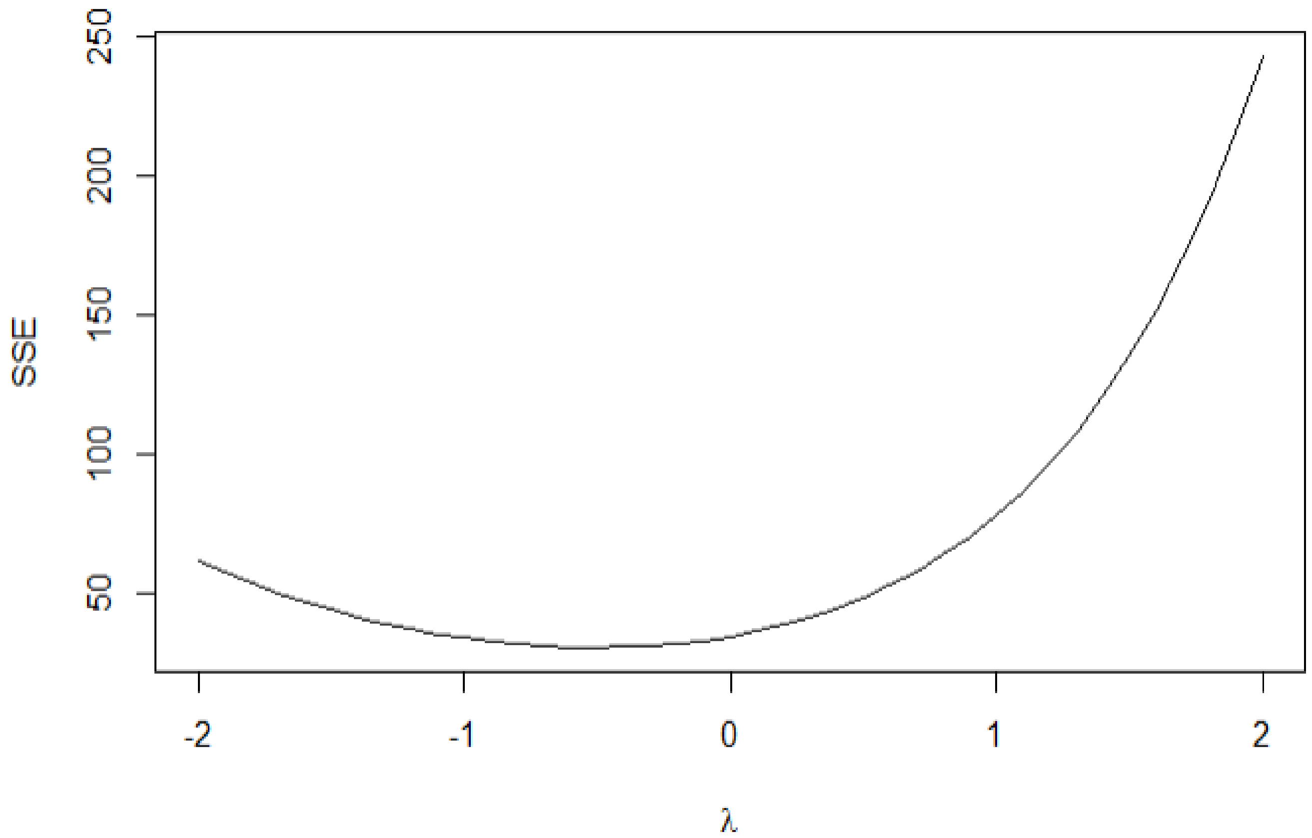
$$\lambda = -0.5 \quad Y' = \frac{1}{\sqrt{Y}}$$

The best $\lambda = -0.515$ $(biggest\ log - likelihood)$




```
library(MASS)
bcmle<-boxcox(lm(plasma~age,data=orig),lambda=seq(-3,3, by=0.1))
lambda<-bcmle$x[which.max(bcmle$y)]
lambda
```

The best $\lambda = -0.5$ $(smallest\ SSE)$



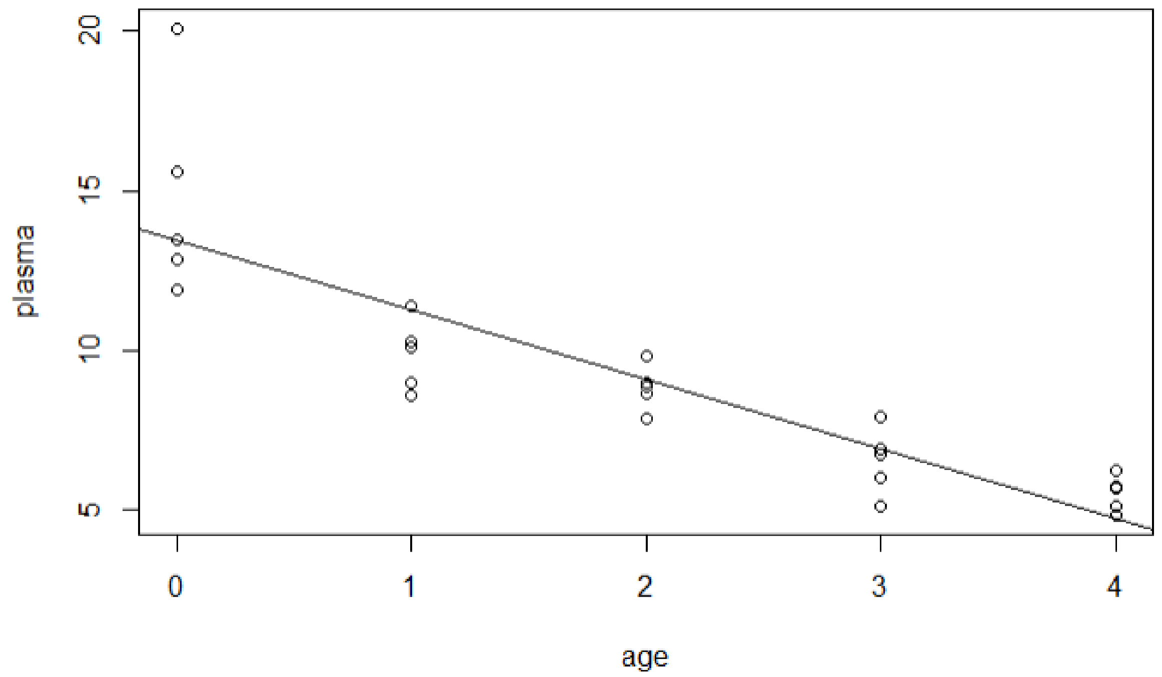

```
library(ALSM)
bcsse<-boxcox.sse(plasma$age,plasma$plasma,l=seq(-2,2,0.1))
lambda<-bcsse$lambda[which.min(bcsse$SSE)]
lambda
```

$Y^{\lambda} = \beta_0 + \beta_1 X, where\ \lambda\ ranges\ from\ -3\ to\ 3, increases\ by\ 0.1)$

Transform $Y' = \frac{1}{\sqrt{Y}}$

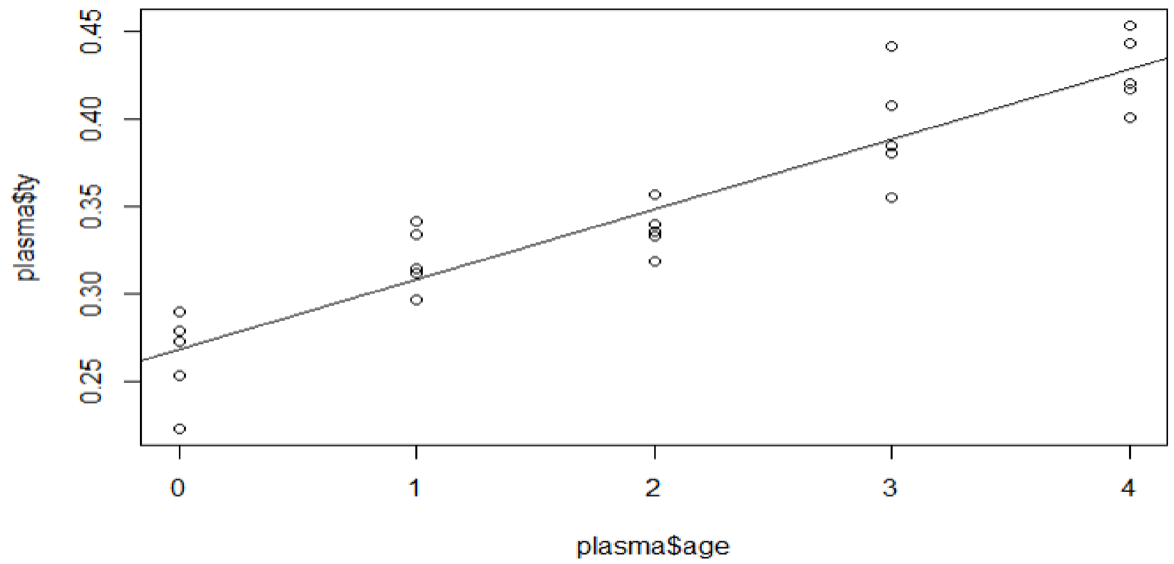
Because s{residual} has the same unit as the response variable, but transformation alters that.

Before




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.4752     0.6379  21.126  < 2e-16 ***
age          -2.1820     0.2604  -8.379 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.841 on 23 degrees of freedom
Multiple R-squared:  0.7532,    Adjusted R-squared:  0.7425 
F-statistic: 70.21 on 1 and 23 DF,  p-value: 1.92e-08
```

After

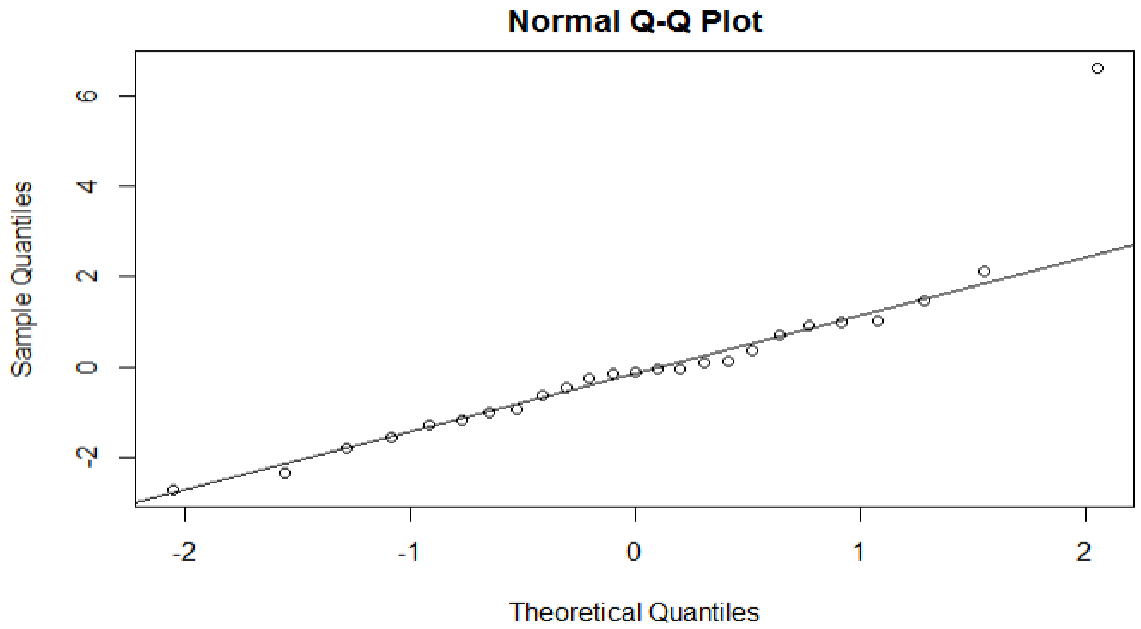


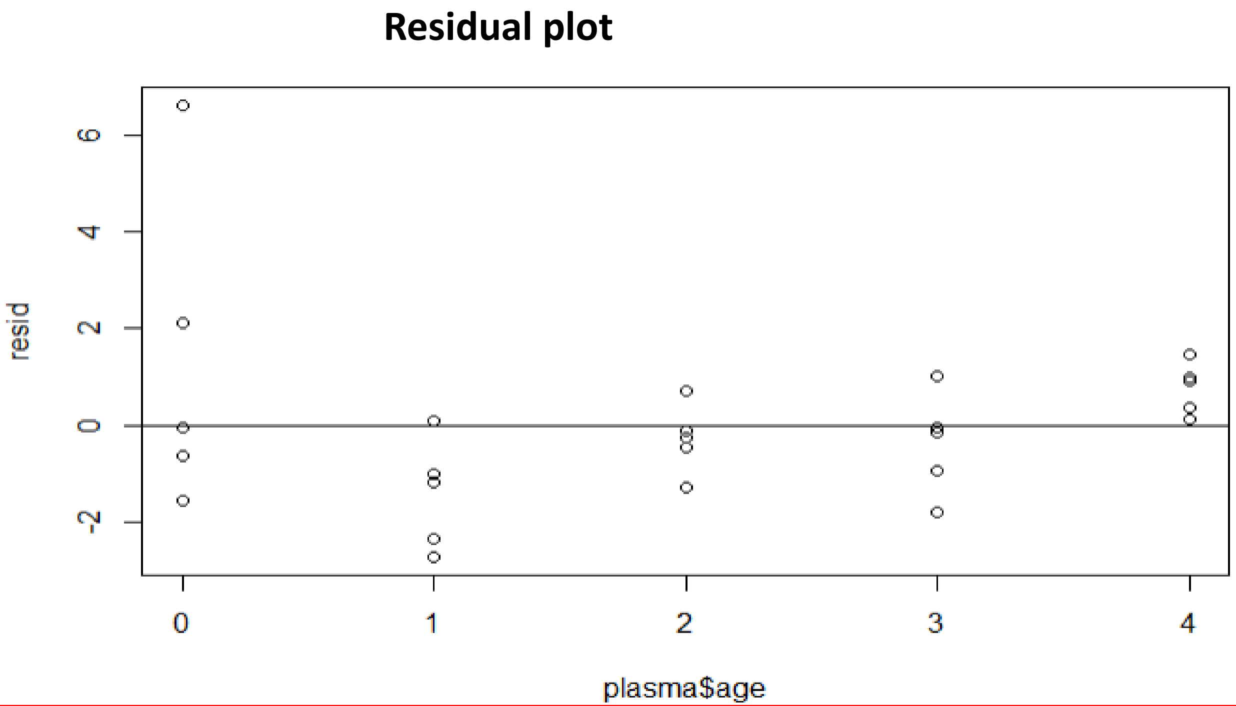

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.268026   0.008033   33.36  < 2e-16 ***
age         0.040062   0.003280   12.22 1.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02319 on 23 degrees of freedom
Multiple R-squared:  0.8665,    Adjusted R-squared:  0.8606 
F-statistic: 149.2 on 1 and 23 DF,  p-value: 1.548e-11
```

Why the s(residual) is not comparable?

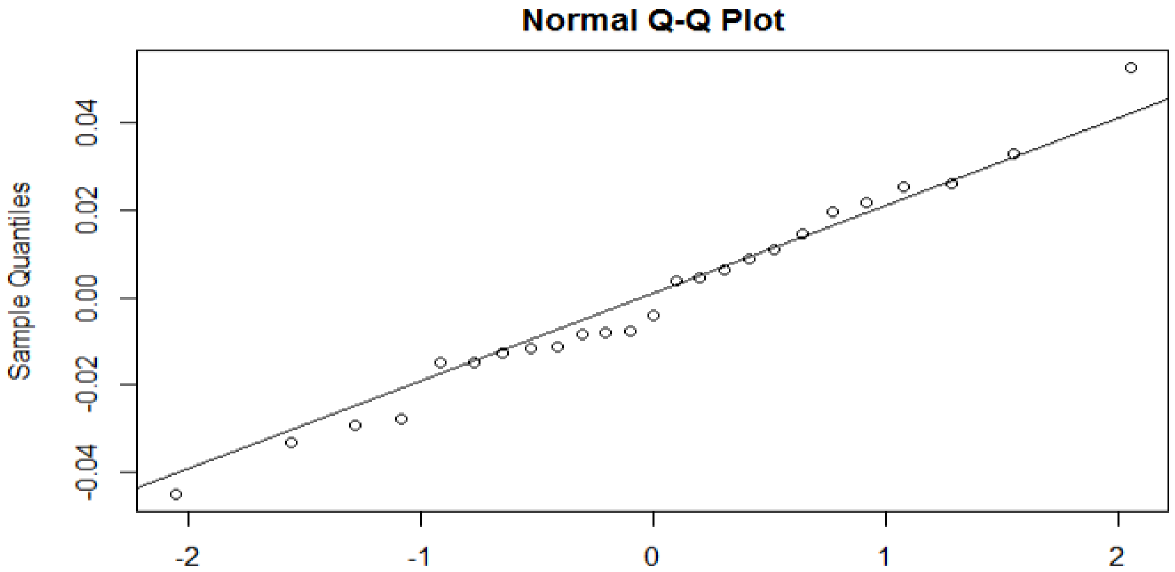
Re-Check Normality and constancy on the residuals

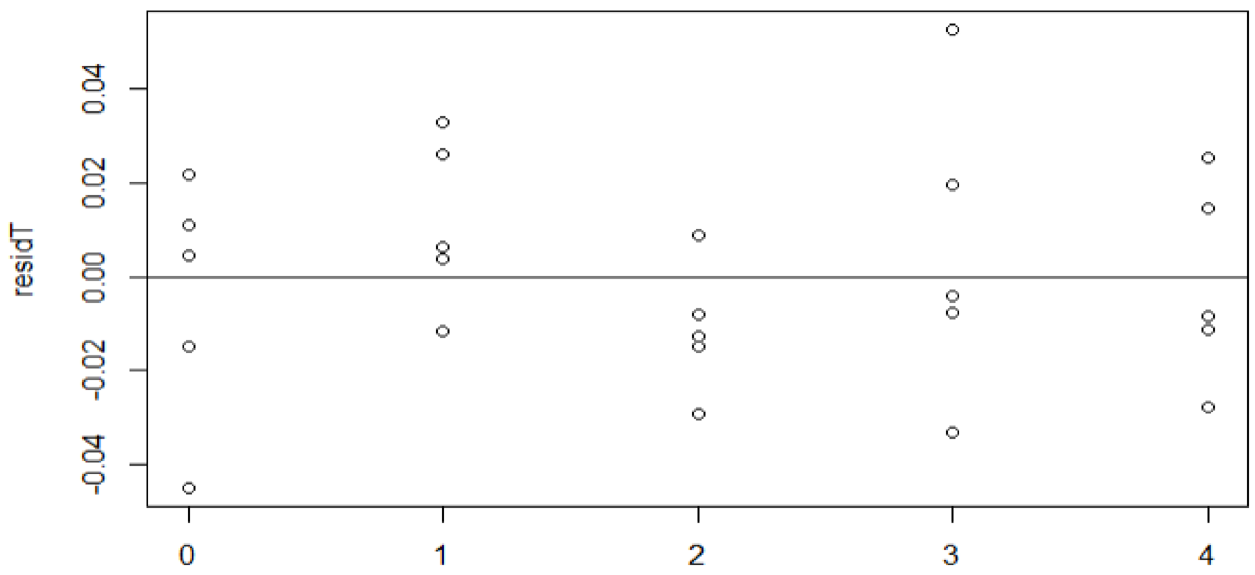
Before





After





Back-transformations

Transformations can improve model performance, but make interpretation hard.

Back transformation lets us make inferences (and **graphs!**) on the original scale.

Very helpful for communicating results to the public.

Interpreting the confidence interval for the mean and single prediction

- In general, let $Y' = f(Y)$and let f' be the **back-transformation function.**

  For example,
  $Y' = f(Y) = Y^2$, the back-transformation function f' does
  $f'(Y') = Y$, so $f'(Y') = \sqrt{Y^2} = Y$

- Then, back transform the mean and single response confidence interval (a, b) as following

  $(f'(a), f'(b))$, For example, $(\sqrt{a}, \sqrt{b})$

- Back transforming the coefficients or the standard error is not accurate.

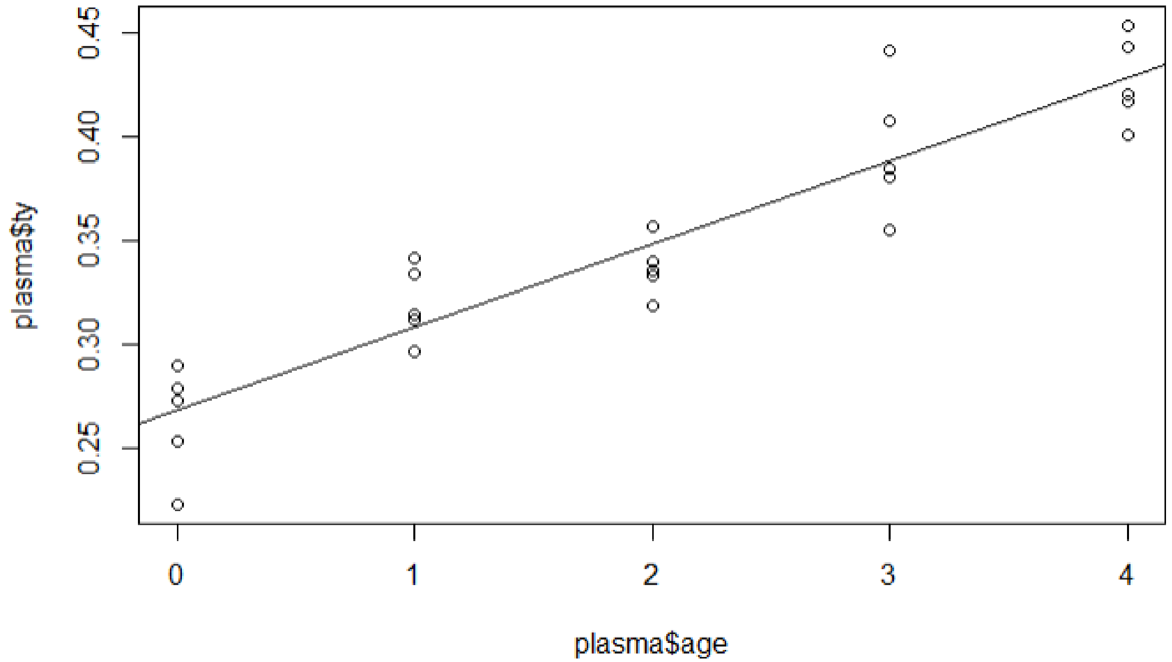
  For example, do not back transform the point estimate with $\hat{Y} = f'(b_0) + f'(b_1)\mathrm{X} = \sqrt{b_0} + \sqrt{b_1}X$,
  $\text{in stead, do } \sqrt{b_0 + b_1 X}$

- If only X is transformed to X', then no need to back transform Y’s estimation because Y hasn’t been transformed.

  For example, $\hat{Y} = b_o + b_1 X'$

Back-transform $Y' = \frac{1}{\sqrt{Y}}$

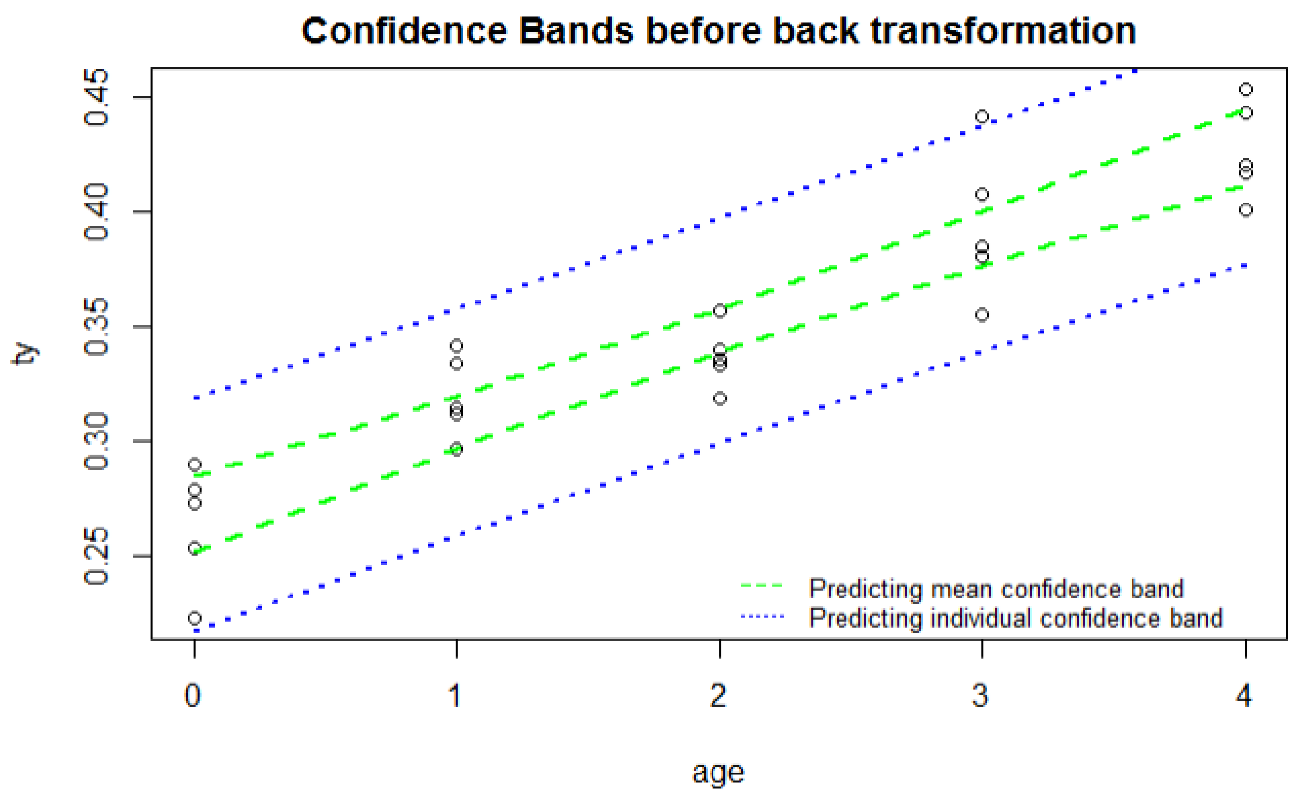
1. The back transform function $f' = \frac{1}{Y'^2} = (Y')^{-2}$
2. The predicted value should be $\hat{Y} = \left(\hat{Y}'\right)^{-2}$
3. The confidence interval for the prediction, either for the mean or single response, should also be back transformed with $(value)^{-2}$.

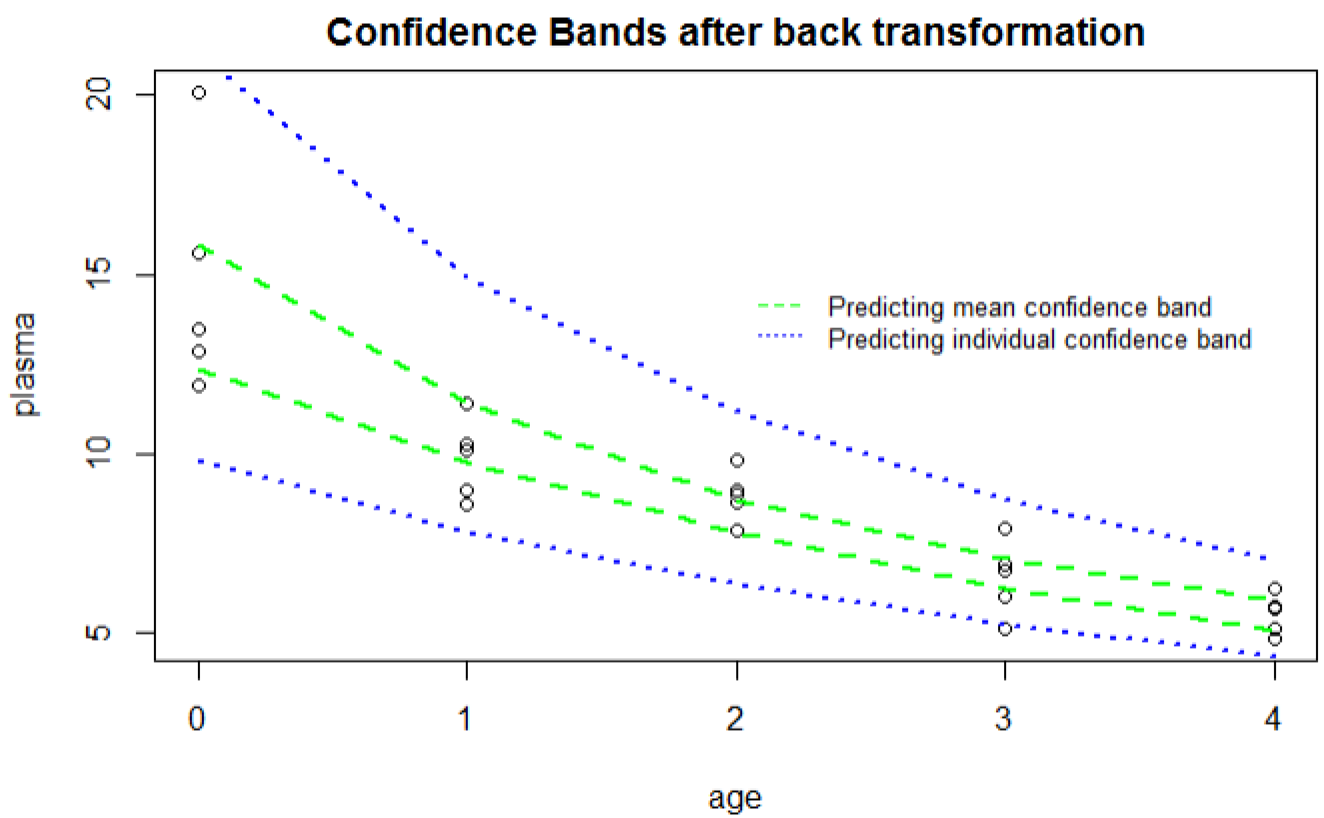



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.268026   0.008033   33.36  < 2e-16 ***
age         0.040062   0.003280   12.22 1.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02319 on 23 degrees of freedom
Multiple R-squared:  0.8665,	Adjusted R-squared:  0.8606 
F-statistic: 149.2 on 1 and 23 DF,  p-value: 1.548e-11
```

$$\frac{1}{\sqrt{Y}} = 0.268 + 0.04(X)$$

Back-transform $Y' = \frac{1}{\sqrt{Y}}$, then $Y = (Y')^{-2}$






```
plot(ty ~ age, plasma, main="Confidence Bands before back transformation")

lines(cim$age, cim$Lower.Band,col="green", lwd=2, lty=2)
lines(cim$age, cim$Upper.Band, col="green", lwd=2, lty=2)
lines(cin$age, cin$Lower.Band,col="blue", lwd=2, lty=3)
lines(cin$age, cin$Upper.Band, col="blue", lwd=2, lty=3)
```

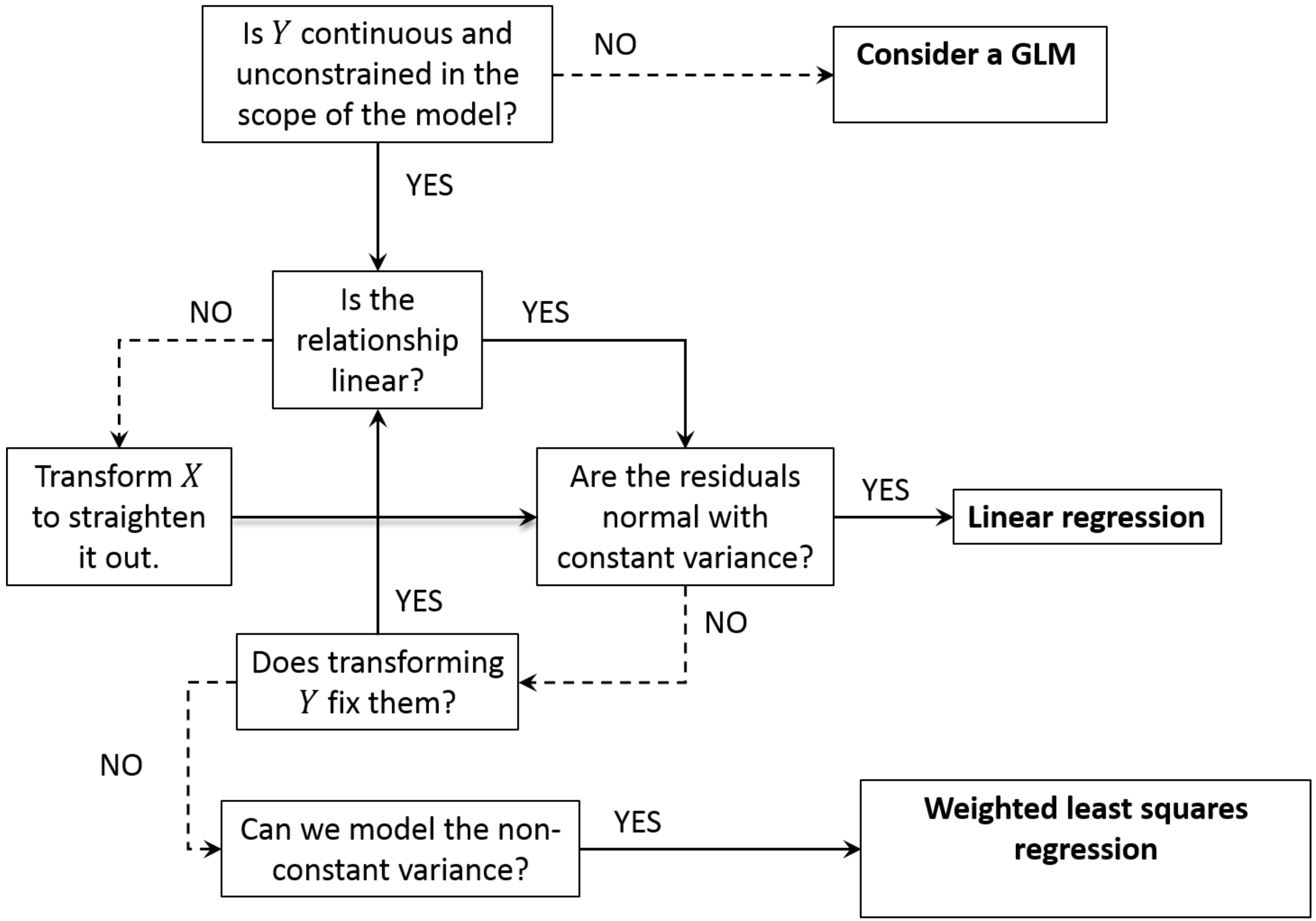
```
plot(plasma ~ age, plasma, main="Confidence Bands after back transformation")

lines(cim$age, (cim$Lower.Band)^(-2),col="green", lwd=2, lty=2)
lines(cim$age, (cim$Upper.Band)^(-2), col="green", lwd=2, lty=2)
lines(cin$age, (cin$Lower.Band)^(-2),col="blue", lwd=2, lty=3)
lines(cin$age, (cin$Upper.Band)^(-2), col="blue", lwd=2, lty=3)
```

Summary of remedial measures

- For nonlinear functional relationships with well behaved residuals
  - Try transforming X
  - May require a polynomial or piecewise fit (we will cover these later)
- For non-constant or non-normal variance, possibly with a nonlinear functional form
  - Try transforming Y
  - The Box-Cox procedure may be helpful
  - If the transformation on Y doesn't fix the non constant variance problem, weighted least squares can be used (we will cover this later).

- Transformations of X and Y can be used together.
- Any time you consider a transformation
  - Remember to recheck all the diagnostics.
  - Consider whether you gain enough to justify losing interpretability.
  - Reciprocal transformations make interpretation especially hard.
  - Consider back-transforming the results of the final model for presentation.
- For very non-normal errors, especially those arising from discrete responses, generalized linear models are often a better option, but linear regression may be “good enough.”

Transformation – our primary tool to improve model fit



Always repeat diagnostic process after transformation