Huiwen Chen ID: 02156341 uniquename:huiwenc

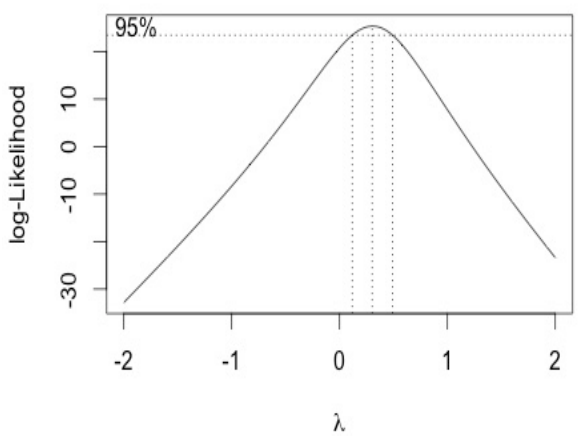
STAT 500 HW6

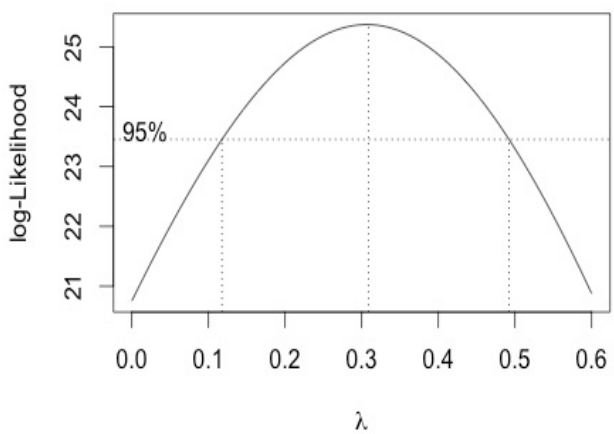
Using the trees data, fit a model with Volume as the response and Girth and Height as predictors.

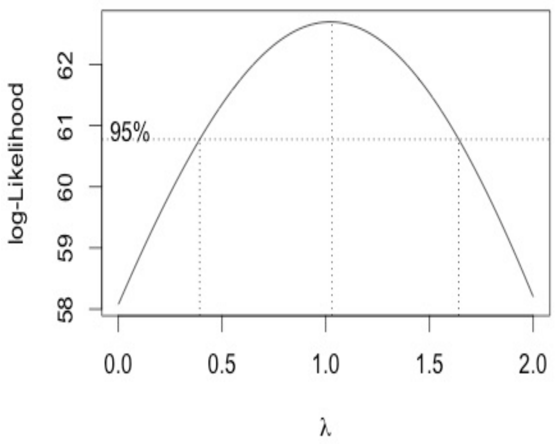
1. Use the Box-Cox method to determine the best transformation on the response. Compare the fits with and without the transformation.

Code:

```
> g1 <-   lm(Volume~ Girth +   Height)

> boxcox(g1, plotit =T)

> boxcox(g1, lambda = seq(0.0, 0.6, by =0.05))

> g2 <- lm(I(Volume^0.3)~Girth + Height)

> summary(g1)

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept) -57.9877       8.6382   -6.713 2.75e-07

Girth        4.7082        0.2643   17.816   < 2e-16

Height       0.3393        0.1302    2.607    0.0145

Residual standard error: 3.882 on 28 degrees of freedom

Multiple R-squared:   0.948,Adjusted R-squared: 0.9442

F-statistic: 255 on 2 and 28 DF,   p-value: < 2.2e-16

boxcox(g2, plotit =T,lambda = seq(0.0, 2, by =0.05))

> summary(g2)

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.194613     0.148552     1.310      0.201

Girth    0.121559     0.004545    26.748     < 2e-16

Height 0.011799     0.002238      5.272     1.32e-05

Residual standard error: 0.06676 on 28 degrees of freedom

Multiple R-squared:   0.9775,     Adjusted R-squared:   0.9759

F-statistic: 609.1 on 2 and 28 DF,   p-value: < 2.2e-16
```







After using Box_Cox method to determine the transformation on the response, it seems like that a cubic root transformation will make a better fit here. Also the new $\hat{\lambda}$ is close to 1 after making cubic transformation on response. Compare with the fit without transformation, in the fit after transformation the predictor Height becomes more significant and R-squared increases slightly. So the fit after transformations is more acceptable.

2. Try adding higher order polynomial terms in the predictors to the original linear model. Comment on the changes in fit.

Code:

```
> g3 <- lm(Volume~ Girth+Height+I(Girth*Height)+I(Girth^2)+I(Height^2))

> summary(g3)

Coefficients:

                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.60706   62.90855   0.105   0.9172
Girth             -5.12160    2.46674  -2.076   0.0483
Height             0.29491    1.77852   0.166   0.8696
I(Girth * Height)  0.06628    0.05671   1.169   0.2535
I(Girth^2)         0.16393    0.10089   1.625   0.1167
I(Height^2)       -0.00494    0.01312  -0.376   0.7097

Residual standard error: 2.655 on 25 degrees of freedom

Multiple R-squared:  0.9783,  Adjusted R-squared:  0.9739

F-statistic:   225 on 5 and 25 DF,  p-value: < 2.2e-16

> ##remove Height^2
```

```
> g4 <- lm(Volume~ Girth+Height+I(Girth*Height)+I(Girth^2))

> summary(g4)

Coefficients:

                   Estimate Std. Error t value   Pr(>|t|)
(Intercept)       26.48906   33.61492   0.788    0.4378
Girth             -4.58977    1.98854  -2.308    0.0292
Height            -0.32992    0.62857  -0.525    0.6041
I(Girth * Height)  0.0570    0.05024   1.135    0.2668
I(Girth^2          0.17071   0.09762   1.749    0.0921 .

Residual standard error: 2.611 on 26 degrees of freedom

Multiple R-squared:  0.9781,  Adjusted R-squared:  0.9748

F-statistic: 290.8 on 4 and 26 DF,  p-value: < 2.2e-16

> ##remove Girth^2

> g5 <- lm(Volume~ Girth+Height+I(Girth*Height))

> summary(g5)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	69.39632	23.83575	2.911	0.00713
Girth	5.85585	1.92134	-3.048	0.00511
Height	-1.29708	0.30984	-4.186	0.00027
I(Girth * Height)	0.13465	0.02438	5.524	7.48e-06

Residual standard error: 2.709 on 27 degrees of freedom

Multiple R-squared: 0.9756, Adjusted R-squared: 0.9728

F-statistic: 359.3 on 3 and 27 DF, p-value: < 2.2e-16

After adding all of the quadratic and linear terms of predictor Girth and Height， we find the Height^2 is the highest order term with the largest p-value, so we remove Height^2 and refit. Then remove Girth^2 similarly, and refit again, as a result all of the predictors left are significant with much smaller p-value than before, but the R-squared dropped slightly in the removing process.

3.Try to improve on the best model from (a) by adding higher order polynomial terms in the predictors to the model. Comment on the changes in fit.

```
> g6 <- lm(I(Volume^0.3)~ Girth+Height+I(Girth*Height)+I(Girth^2)+I(Height^2))

> summary(g6)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.4693099	1.6587938	-0.283	0.7796	
Girth	0.1156295	0.0650440	1.778	0.0876	.
Height	0.0305242	0.0468965	0.651	0.5211	
I(Girth * Height)	0.0004603	0.0014953	0.308	0.7608	
I(Girth^2)	-0.0010635	0.0026603	-0.400	0.6927	
I(Height^2)	-0.0001637	0.0003460	-0.473	0.6403	

Residual standard error: 0.07001 on 25 degrees of freedom

Multiple R-squared: 0.9779, Adjusted R-squared: 0.9735

F-statistic: 221.6 on 5 and 25 DF, p-value: < 2.2e-16

```
> ## remove Girth^2

> g7 <- lm(I(Volume^0.3)~ Girth+Height+I(Girth*Height)+I(Height^2))

> summary(g7)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6288141	1.5838634	-0.397	0.6946
Girth	0.1259222	0.0587580	2.143	0.0416
Height	0.0332372	0.0456470	0.728	0.4730
I(Girth * Height)	-0.0000543	0.0007483	-0.073	0.9427
I(Height^2)	-0.0001390	0.0003349	-0.415	0.6815

Residual standard error: 0.06887 on 26 degrees of freedom

Multiple R-squared: 0.9778, Adjusted R-squared: 0.9744

F-statistic: 286.2 on 4 and 26 DF, p-value: < 2.2e-16

```
> ##remove Height^2

> g8 <- lm(I(Volume^0.3)~ Girth+Height+I(Girth*Height))

> summary(g8)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.0214129	0.5967006	-0.036	0.97164	
Girth	0.1394738	0.0480984	2.900	0.00734	
Height	0.0145742	0.0077566	1.879	0.07109	.
I(Girth * Height)	-0.0002284	0.0006103	-0.374	0.71118	

---Residual standard error: 0.06781 on 27 degrees of freedom

Multiple R-squared: 0.9776, Adjusted R-squared: 0.9752

F-statistic: 393.6 on 3 and 27 DF, p-value: < 2.2e-16

```
> ##remove Girth*Height
> g9 <- lm(I(Volume^0.3)~ Girth+Height)
> summary(g9)
 Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 0.194613   0.148552   1.310      0.201
Girth   0.121559     0.004545   26.748    < 2e-16
Height 0.011799     0.002238     5.272   1.32e-05
```

Residual standard error: 0.06676 on 28 degrees of freedom

Multiple R-squared: 0.9775, Adjusted R-squared: 0.9759

F-statistic: 609.1 on 2 and 28 DF, p-value: < 2.2e-16

After adding higher order polynomial terms to the best model, we find almost all of the predictors are not significant, then we remove the most non-significant and highest order term Girth^2. The p-values of left predictors change a little, and Height^2 and Girth*Height are still extremely unsignificant, so we remove Height^2 at first and refit, then find Girth* Height becomes the only unsignificant predictor. So we remove it and all the predictors left are significant. The R-squared is more stable in this removing process. The final model is a linear one.