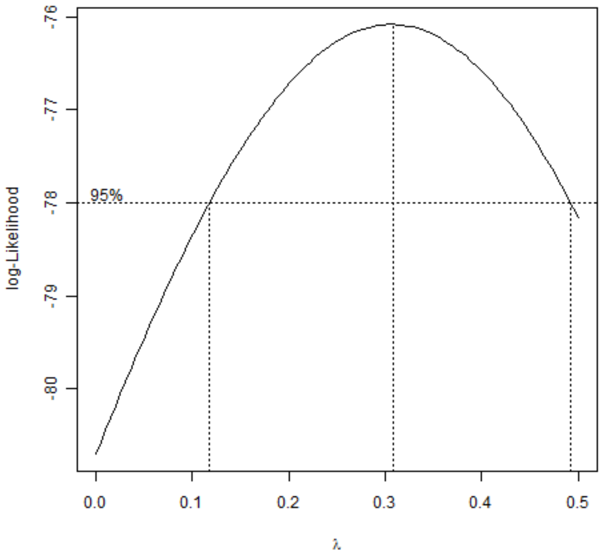
Stat 500 – Homework 6 (Solutions)

(a) Read in data, fit the model and transform it.

```
>library(MASS)
>library(faraway)
>data(trees)
>a=boxcox(lm(Volume~Girth+Height,data=trees),lambda=seq(0,.5,by=.05))
```



A cube-root or fourth-root transformation ($\lambda = 0.33$ or 0.25) would work fine according to the CI from the Box-Cox method. Here we will work with $\lambda = 1/3$ since cube root is easily interpretable as the data is volume. Below is the original fit before transformation:

```
>g1=lm(Volume~Girth+Height,data=trees)
>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
```

Next is the fit after transformation:

```
>g2=lm(I(Volume^(1/3))~Girth+Height,data=trees)
>summary(g2)

Coefficients:
              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  -0.054544    0.180435   -0.302     0.765
Girth         0.148286    0.005520   26.864    < 2e-16 ***
Height        0.014186    0.002719    5.218   1.53e-05 ***

Residual standard error: 0.08108 on 28 degrees of freedom
Multiple R-Squared: 0.9776,     Adjusted R-squared: 0.9761
F-statistic: 612.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

Note that R^2 has increased in the transformed model, which indicates a better overall fit.

(b) In order to keep the model simpler we do not go beyond quadratic terms to improve the model. We start with the following model:

```
>g3=lm(Volume~Girth+Height+I(Girth^2)+I(Height^2)+Girth*Height,data=trees)
>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^2)      0.16393    0.10089     1.625     0.1167
I(Height^2)    -0.00494    0.01312    -0.376     0.7097
Girth:Height    0.06628    0.05671     1.169     0.2535

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
```

We do backward elimination on the above model with the restriction that no first order term is removed until all second order terms involving that variable are removed and we retain only the significant predictors. We get the following as the final model:

```
Coefficients:
              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  -9.92041     10.07911    -0.984     0.333729
Girth        -2.88508      1.30985    -2.203     0.036343 *
Height        0.37639      0.08823     4.266     0.000218 ***
I(Girth^2)    0.26862      0.04590     5.852     3.13e-06 ***

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,     Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
```

We see that the R^2 value of this model and the transformed model in part (a) are almost same. So we can conclude that transforming the response and adding a quadratic term in one of the predictor gives almost the same improvement in the fit.

(c) Now we add the same polynomial terms in the predictors to the transformed model in part (a):

```
>g4=lm(I(Volume^(0.33))~Girth+Height+I(Girth^2)+I(Height^2)+Girth*Height,data=trees)
>summary(g4)

Coefficients:
                   Estimate    Std. Error   t value  Pr(>|t|)
(Intercept)       -0.7382295   2.0217696    -0.365    0.718
Girth              0.1270359   0.0792768     1.602    0.122
Height             0.0360557   0.0571584     0.631    0.534
I(Girth^2)        -0.0010024   0.0032425    -0.309    0.760
I(Height^2)       -0.0001988   0.0004217    -0.471    0.641
Girth:Height       0.0006339   0.0018225     0.348    0.731

Residual standard error: 0.08533 on 25 degrees of freedom
Multiple R-Squared: 0.9779,     Adjusted R-squared: 0.9735
F-statistic: 221.2 on 5 and 25 DF,  p-value: < 2.2e-16
```

We do similar backward elimination on the above model as in part (b). We get the following as the final model:

```
Coefficients:
             Estimate    Std. Error  t value   Pr(>|t|)
(Intercept)  -0.054544    0.180435   -0.302     0.765
Girth         0.148286    0.005520   26.864    < 2e-16 ***
Height        0.014186    0.002719   5.218     1.53e-05 ***

Residual standard error: 0.08108 on 28 degrees of freedom
Multiple R-Squared: 0.9776,     Adjusted R-squared: 0.9761
F-statistic: 612.4 on 2 and 28 DF, p-value: < 2.2e-16
```

Thus, we can infer that transforming the response and adding quadratic terms in predictors are not both necessary. We can use either of the two models,

```
Model 1: Volume^(1/3) ~ Girth + Height
Model 2: Volume ~ Girth + Height + I(Girth^2)
```

Though Model 1 is perhaps more easily interpretable than Model 2 (since cube-root of volume is in the same units as length).