STATS 500 HOMEWORK 7 YUANYIN

Problem 1

Before we take different method to fit the model, we first look at the original linear model and its results:

```
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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

Both coefficients are significant and $R^2 = 0.948$ which is very close to 1, the original fits very well already, however, we still want to find out if we can improve the fit.

(a) First we use Box-Cox method to determine the best transformation on the response. Compute λ and the 95% confidence interval is as follows:

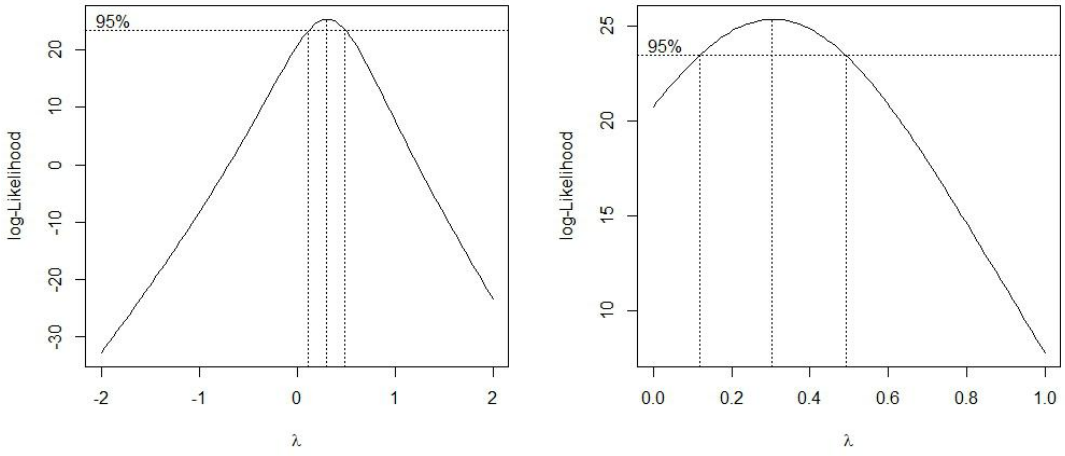

Fig 1 CI of λ

We take a reasonable value of λ in the CI above: $\lambda = 1/3 \approx 0.33$, and we refit the model with transformation $y \to y^{\lambda}$, the result is as follows:

```
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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

Finding that the two coefficient became much more significant than the model before (because the all the p-value become smaller), besides, the new $R^2 = 0.9776$ is even larger than before, which means the model fits better than the original model. Our final mode is as follows:

$$y^{1/3} = \beta_0 + \beta_{Girth} x_{Girth} + \beta_{Height} x_{Height}$$

(b) Next we use another method to improve the model which is adding higher order polynomial terms in the predictors. We use backward elimination to determine our final model and the result is as follows:

```
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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

In our new model, we add new term I(Girth^2). We found that the p-value of all parameters in new model are still less than 0.05, which means they are all significant, and we found $R^2 = 0.9771$ in our new model which is higher than the original model but rather close to model in Box-Cox mehod, besides, the parameter of Girth and Height isn't that significant than the parameters in the model of Box-Cox method (compare p-value of two model). What's more, the residual standard error in this model is 2.625 which is much larger than the one in Box-Cox method which is 0.081. In conclusion, this model fits better than the original model but not better than model in (a). Our final model is as follows:

$$y = \beta_0 + \beta_{Girth} x_{Girth} + \beta_{Height} x_{Height} + \beta_{Girth} x_{Girth}^2$$

(c) Now we want to see if the model fits better if we use adding higher order polynomial terms method on the model we get in (a). Utilize testing-based backward elimination and what we get in the end is still the model that we get in (a), which is the regression of "Volume^(.33) ~ Girth + Height, data = trees" (i.e. $y^{1/3} = \beta_0 + \beta_{Girth} x_{Girth} + \beta_{Height} x_{Height}$). In conclusion, there is no change of fit comparing with model in (a).

First look at the model we get in (a) and (c), which is $y^{1/3} = \beta_0 + \beta_{Girth} x_{Girth} + \beta_{Height} x_{Height}$.

We check its residuals to see if there is nonlinearity of our model, the plot of residuals vs fitted values is as follows:

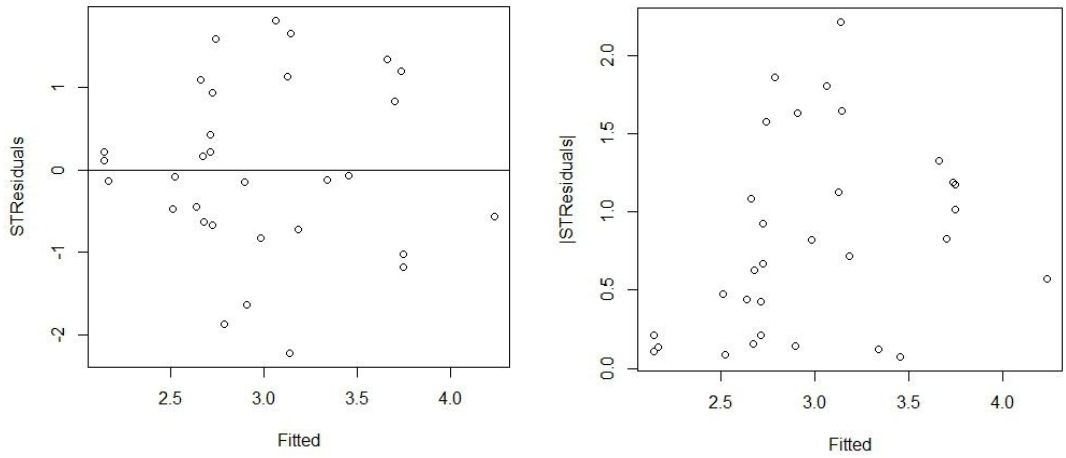

Fig 2 residuals vs fitted values (left) & |residuals| vs fitted values (right)

We can see that the distribution of residuals look like randomly and evenly, there seems no nonlinearity of our model.

Then we check the normality of errors of the model, using QQ-plot:

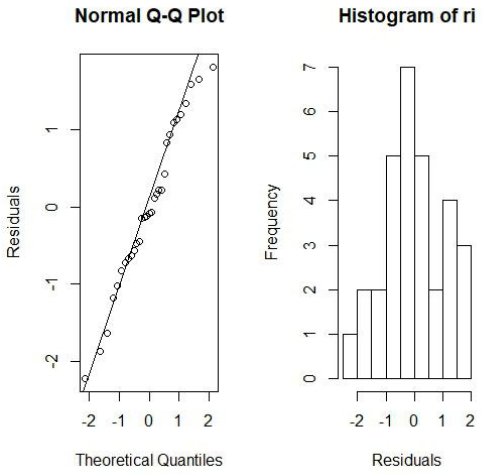

Fig 3 QQ-plot of model (a)

We can see that the points are all very close to the line and we can assume that the errors obey the normal distribution.

For outliers, we compute the p-value is 0.03503326, comparing with adjusted $\alpha = 0.05/31 = 0.001612903$, p-value is larger than α which means we can't reject the null hyphothesis. Thus, there is no outlier in this model.

Next we want to make diagnostics on model in (b), from problem (b) above we know the model of (b) is $y = \beta_0 + \beta_{Girth} x_{Girth} + \beta_{Height} x_{Height} + \beta_{Girth} x_{Girth}^2$.

First we check nonlinearity of our model, the plot of residuals vs fitted values is as follows:

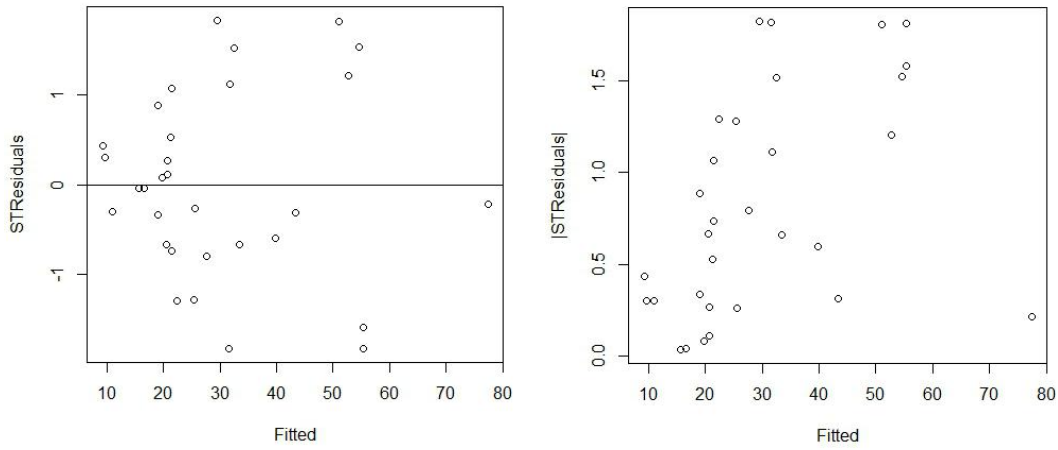

Fig 4 residuals vs fitted values (left) & |residuals| vs fitted values (right)

It seems like there is still nothing wrong with the nonlinearity. So we can hold the nonlinearity assumption.

Then we check the normality of errors of the model, using QQ-plot:

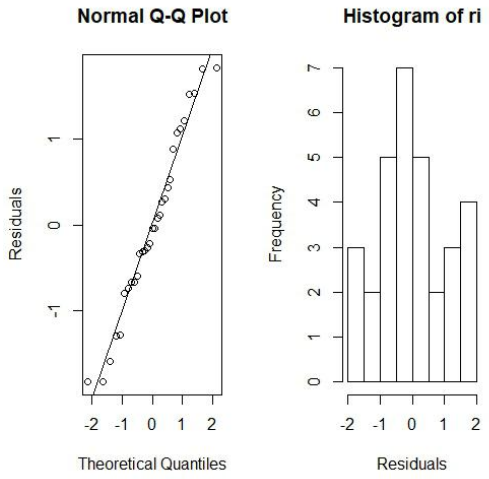

Fig 5 QQ-plot of model (b)

We can see that the points are all very close to the line and we can assume that the errors obey the normal distribution.

For outliers, we compute the p-value is 0.07927363, comparing with adjusted $\alpha = 0.05/31 = 0.001612903$, p-value is larger than α which means we can't reject the null hyphothesis. Thus, there is no outlier.

Appendix

```
library(faraway)
data(trees)
library(MASS)
# Box-Cox method for trees data
y = lm(Volume ~ Girth + Height, data = trees)
summary(y)
boxcox(y, plotit = T)
```

```
boxcox(y, plotit = T, lambda = seq(0, 1, by = 0.1))
ybc = lm(Volume^(.33) ~ Girth + Height, data = trees)
summary(ybc)
# polynomials backward elimination
## 2nd degree
summary(lm(Volume ~ Girth + Height + I(Girth^2) + I(Height^2) + I(Girth * Height),
data = trees))
summary(lm(Volume ~ Girth + Height + I(Girth^2) + I(Girth * Height), data = trees))
summary(lm(Volume ~ Girth + Height + I(Girth^2), data = trees))
ypol = lm(Volume ~ Girth + Height + I(Girth^2), data = trees)
# box-cox & polynomials backward elimination
summary(lm(Volume^(.33) ~ Girth + Height + I(Girth^2) + I(Height^2) + I(Girth
* Height), data = trees))
summary(lm(Volume^(.33) ~ Girth + Height + I(Height^2) + I(Girth * Height), data
= trees))
summary(lm(Volume^(.33) ~ Girth + Height + I(Height^2), data = trees))
summary(lm(Volume^(.33) ~ Girth + Height, data = trees))
# for (a) and (c) model
## nonlinearlity
ri = rstudent(ybc)
plot(ybc$fitted, ri, xlab = "Fitted", ylab = "STResiduals")
abline(h = 0)
## QQ plot
par(mfrow = c(1,2))
qqnorm(ri, ylab = "Residuals")
qqline(ri)
## Histogram
hist(ri, xlab = "Residuals")
max(abs(ri))
which(ri == max(-abs(ri)))
## compute p-value
2*(1 - pt(max(abs(ri)), df = 31-3-1))
## compare to alpha/n
0.05/31
# for (b) model
## nonlinearlity
ri = rstudent(ypol)
plot(ypol$fitted, ri, xlab = "Fitted", ylab = "STResiduals")
abline(h = 0)
## QQ plot
par(mfrow = c(1,2))
qqnorm(ri, ylab = "Residuals")
qqline(ri)
## Histogram
```

```
hist(ri, xlab = "Residuals")
max(abs(ri))
which(ri == max(-abs(ri)))
## compute p-value
2*(1 - pt(max(abs(ri)), df = 31-4-1))
## compare to alpha/n
0.05/31
```