HW6

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

Call:

lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:

Min 1Q Median 3Q Max

-6.4065 -2.6493 -0.2876 2.2003 8.4847

The model is kind of well fitted, the adjusted R-squared reaches 0.944. And the p-value indicates significance for the model.

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

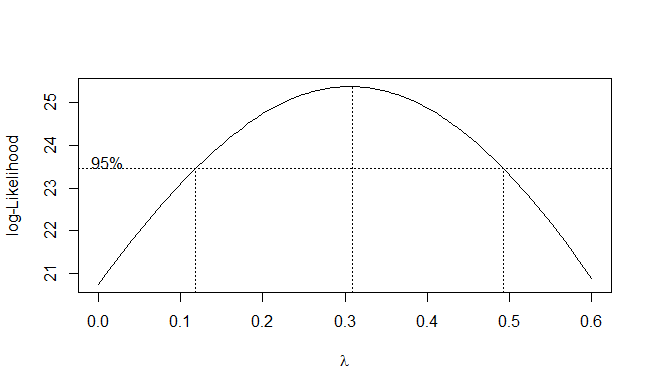
Residual standard error: 3.882 on 28 degrees of freedom

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

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

(a)

Use the Box-Cox method to determine the best transformation on the response.

From the Box-Cox plot we can see that the mean of lambda is close to 0.31.

Choose lambda=0.31 to make a transformation for the response.

Then we get a better model:

Call:

lm(formula = v1 ~ Girth + Height, data = trees)

Residuals:

Transformation:

Min 1Q Median 3Q Max

-0.43760 -0.14185 -0.01797 0.19188 0.37596

Coefficients:

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

(Intercept) -2.848296 0.511617 -5.567 5.90e-06 \*\*\*

The transformation brings a better fit. It brings more significance for Height and a larger adjusted R-squared.

Girth 0.419402 0.015652 26.796 < 2e-16 \*\*\*

Height 0.040513 0.007708 5.256 1.38e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2299 on 28 degrees of freedom

Multiple R-squared: 0.9776, Adjusted R-squared: 0.976

F-statistic: 610.6 on 2 and 28 DF, p-value: < 2.2e-16

(b)

Add Girth^2,Height^2,Height\*Girth to the original linear model.

We get the following estimate

Call:

lm(formula = Volume ~ Girth + Height + I(Girth^2) + I(Height^2) +

I(Girth \* Height), data = trees)

Residuals:

Min 1Q Median 3Q Max

-4.7174 -0.8116 -0.0334 1.8299 3.9870

Coefficients:

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

While Girth stays significant, however Height and quadratic terms are not significant.

(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

I(Girth \* Height) 0.06628 0.05671 1.169 0.2535

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Now we use backward elimination:

First we remove the Height2 with the largest p-value.

Call:

lm(formula = Volume ~ Girth + Height + I(Girth^2) + I(Girth \*

Height), data = trees)

Residuals:

Min 1Q Median 3Q Max

-5.0748 -0.8494 0.0051 1.8396 4.0604

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 \*

We can see that Girth2 has trend to be significant and the fit is slightly better.

Height -0.32992 0.62857 -0.525 0.6041

I(Girth^2) 0.17071 0.09762 1.749 0.0921 .

I(Girth \* Height) 0.05701 0.05024 1.135 0.2668

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Now we remove the Girth\*Height

Call:

lm(formula = Volume ~ Girth + Height + I(Girth^2), data = trees)

Residuals:

Min 1Q Median 3Q Max

-4.2928 -1.6693 -0.1018 1.7851 4.3489

Coefficients:

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

(Intercept) -9.92041 10.07911 -0.984 0.333729

This time all the predictors reach significant and the fit of the model is better than the original model.

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

(c)

Add Girth^2,Height^2,Height\*Girth to the Transformed linear model.

Call:

lm(formula = v1 ~ Girth + Height + I(Girth^2) + I(Height^2) +

I(Girth \* Height), data = trees)

Residuals:

Min 1Q Median 3Q Max

-0.4306 -0.1146 -0.0097 0.1911 0.3910

Coefficients:

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

(Intercept) -5.019781 5.721525 -0.877 0.3887

Significance of Girth and Height is apparently decreased and the fit is not as good as the transformed model.

Girth 0.385700 0.224350 1.719 0.0979 .

Height 0.104201 0.161756 0.644 0.5253

I(Girth^2) -0.003391 0.009176 -0.370 0.7148

I(Height^2) -0.000564 0.001193 -0.473 0.6406

I(Girth \* Height) 0.001657 0.005158 0.321 0.7507

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2415 on 25 degrees of freedom

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

F-statistic: 221.5 on 5 and 25 DF, p-value: < 2.2e-16

Now we use backward elimination:

First we remove the Girth\*Height with the largest p-value.

Call:

lm(formula = v1 ~ Girth + Height + I(Girth^2) + I(Height^2),

data = trees)

Residuals:

Min 1Q Median 3Q Max

-0.42037 -0.13217 -0.03077 0.19188 0.39529

Coefficients:

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

(Intercept) -5.2088124 5.5921641 -0.931 0.36019

Slightly better with Girth significant.

Girth 0.4438187 0.1303382 3.405 0.00216 \*\*

Height 0.0998130 0.1583739 0.630 0.53404

I(Girth^2) -0.0008533 0.0045872 -0.186 0.85387

I(Height^2) -0.0003976 0.0010565 -0.376 0.70975

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2373 on 26 degrees of freedom

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

F-statistic: 286.7 on 4 and 26 DF, p-value: < 2.2e-16

Now we remove the Girth2 with the largest p-value.

Call:

lm(formula = v1 ~ Girth + Height + I(Height^2), data = trees)

Residuals:

Min 1Q Median 3Q Max

-0.41867 -0.13260 -0.02974 0.19122 0.40716

Coefficients:

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

(Intercept) -5.5212304 5.2377710 -1.054 0.301

Only Girth stays significant.

Girth 0.4197608 0.0158772 26.438 <2e-16 \*\*\*

Height 0.1124482 0.1404872 0.800 0.430

I(Height^2) -0.0004813 0.0009385 -0.513 0.612

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.233 on 27 degrees of freedom

Multiple R-squared: 0.9778, Adjusted R-squared: 0.9753

F-statistic: 396.4 on 3 and 27 DF, p-value: < 2.2e-16

In summary, the transformed model itself has already got the best fit for all predictors significant.

R-code Used：

library(faraway)

attach(trees)

library(MASS)

g<-lm(Volume~Girth+Height,trees)

boxcox(g,plotit=T,lambda=seq(0,0.6,by=0.1))

trees$v1=(Volume^0.31-1)/0.31

g1<-lm(v1~Girth+Height,trees)

plot(g$fitted,g$residuals,xlab="Fitted",ylab="Residuals",main="Original Model")

plot(g1$fitted,g1$residuals,xlab="Fitted",ylab="Residuals",main="Transformed Model")

g2<-lm(Volume~Girth+Height+I(Girth^2)+I(Height^2)+I(Girth\*Height),trees)

g21<-lm(Volume~Girth+Height+I(Girth^2)+I(Girth\*Height),trees)

g22<-lm(Volume~Girth+Height+I(Girth^2),trees)

g3<-lm(v1~Girth+Height+I(Girth^2)+I(Height^2)+I(Girth\*Height),trees)

g31<-lm(v1~Girth+Height+I(Girth^2)+I(Height^2),trees)

g32<-lm(v1~Girth+Height+I(Height^2),trees)