Anthony Cunningham

STAT 3200

Due 2/24/2017

**Homework 4**

#1. a) > library(car)

> data(trees)

> dim(trees)

[1] 31 3 \*There are 3 variables in “trees” dataset

> names(trees) Their names are “Girth”, “Height”, and

[1] "Girth" "Height" "Volume" “Volume”.

> nrow(trees)

[1] 31 \*There are 31 observations in this dataset.

b) > plot(trees)



> cor(trees)

Girth Height Volume

Girth 1.0000000 0.5192801 0.9671194

Height 0.5192801 1.0000000 0.5982497

Volume 0.9671194 0.5982497 1.0000000

\*It appears that the “Girth” and “Volume” variables have the strongest linear relationship, which can be seen in the correlation matrix, which shows the respective bivariate correlation coefficients (r) of the variable in a given column vs. the variable in a given row. The correlation coefficient is strongest (closest to +/- 1) for the relationship between Girth and Volume, which carries a correlation coefficient of about 0.967.

c) It is safe to assume linearity between Volume and Girth, as the linear relationship is very strong between those two variables. It takes a little more convincing, however, to assume that Volume and Height have a linear relationship. While the linear relationship isn’t very strong, the shape of the relationship doesn’t appear to be curved in any way, so it seems reasonable to assume linearity between Volume and Height.

#2. a) Volume*i* = B0 + B1(Girth)*i* + B2(Height)*i* + *ei*  *e* ~ N(0, o2)

\*Assumptions: 1. Linearity between Volume and predictor variables Girth and Height

2. Constant variance of errors

3. Errors are normally-distributed

4. Errors are independent of each other

b) > attach(trees)

> fit.mlr = lm(Volume ~ Height + Girth)

> summary(fit.mlr)

Coefficients:

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

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

Height 0.3393 0.1302 2.607 0.0145 \*

Girth 4.7082 0.2643 17.816 < 2e-16 \*\*\*

Residual standard error: 3.882 on 28 degrees of freedom

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

\* Volume-hat = -57.988 + 4.708(Girth) + 0.339(Height) + *e* *e* ~ N(0, (3.882)2)

\* R used the Least Squares Method to obtain the model fit.

c) b1=4.708= the estimated change in Volume after a 1 inch increase in Girth, holding all else constant

b2=0.339= the estimated change in Volume after a 1 inch increase in Height, holding all else constant

*o*2-hat=(3.882)2=15.07= the dispersion around Volume at a given (Girth, Height) pair

d) R2=0.948= the proportion of variation in Volume explained by both Girth and Height

Adjusted R2= 1 – (1 - R2)((n – 1)/(n – k – 1)) = 1 – (1- 0.948)((31 – 1)/(31 – 2 – 1)) =

1 – (0.052)(30/28) = 1 – 0.05571 = 0.9442 = Adjusted R2

\* Yes, the calculated adjusted R2 is the exact same as in the R summary output

e) > par(mfrow=c(2,1))

> plot(fit.mlr$residuals)

> abline(h=0)

> qqnorm(fit.mlr$residuals)

> qqline(fit.mlr$residuals)



\* The data appears to spread out more for higher fitted values, so the constant variance assumption may be violated. The QQ Plot, however, doesn’t show drastic departure from normality, with the exception of a little bump for some of the higher fitted values. A transformation of the data could be in order to disperse the errors more evenly and get rid of that bump in the QQ Plot. A log or square-root transformation could be suitable since the variability of the errors tends to increase for larger fitted values. However, one could argue that these assumptions aren’t violated tremendously, and, thus, will not deem transforming the data necessary.

f) 1) H0: BGirth = BHeight = 0

Ha: BGirth ≠ 0 or BHeight ≠ 0 or Both ≠ 0

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

2) Overall F-test is 225

Numerator df = k = 2; Denominator df = n – k – 1 = 28

p-value < 2.2 x 10-16

3) We reject H0, that neither Girth nor Height are significant predictors of Volume.

That is, we think that at least one of the predictors (Girth, Height or both) have some significance on Volume.

g) > RegSS = sum((fit.mlr$fitted.values-mean(Volume))^2)

> RegSS

[1] 7684.163

> RSS = sum((Volume-fit.mlr$fitted.values)^2)

> RSS

[1] 421.9214

> F = (RegSS/2)/(RSS/28)

> F

[1] 254.9723

> pf(F,2,28,lower.tail=FALSE)

[1] 1.071238e-18

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Sum of Squares | df | Mean Square | F |
| Regression  Residuals | 7684.163  421.9214 | 2  28 | 3842.0815  15.0686 | 254.9723 |
| Total | 8106.084 | 30 |  |  |

p-value = 1.07 x 10-18

We reject H0 due to significantly small p-value (i.e. p < 0.05)

h) R2 = RegSS/TSS = 7684.163/8106.084 = 0.948 = R2

\* Yes, the calculated R2 is exactly the same as in the R summary output (Question 2a)

i) > confint(fit.mlr)

2.5 % 97.5 %

(Intercept) -75.68226247 -40.2930554

Height 0.07264863 0.6058538

Girth 4.16683899 5.2494820

j) 1) H0: BHeight = 0

Ha: BHeight ≠ 0

2) t0 = 2.607 \* Values come from summary table in 2a

df = n – k – 1 = 28

p-value = 0.0145

3) We can reject H0. Therefore, Height is determined to be a significant predictor for Volume after accounting for the effect of Girth.

\*Yes, my conclusion is consistent with the 95% CI, as 0 does not fall within the range of possible values of BHeight.

k) 1) H0: BGirth = 0

Ha: BGirth ≠ 0

2) t0 = 17.816

df = n – k – 1 = 28

p-value < 2 x 10-16

3) We can reject H0. Therefore, Girth is determined to be a significant predictor for Volume after accounting for the effect of Height.

\*Yes, my conclusion is consistent with the 95% CI, as 0 does not fall within the range of possible values of BGirth.

#3. a) > fit.slr = lm(Volume ~ Height)

> summary(fit.slr)

Coefficients:

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

(Intercept) -87.1236 29.2731 -2.976 0.005835 \*\*

Height 1.5433 0.3839 4.021 0.000378 \*\*\*

---

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

Residual standard error: 13.4 on 29 degrees of freedom

Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358

F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

b) The estimated slope for predictor Height is interpreted as the estimated change in the Volume of a black cherry tree for every 1 foot increase in the Height of a particular cherry tree.

The estimated coefficient for Height is different in SLR from MLR due to the fact that SLR completely ignores any other predictor (placing it in the error term instead) while MLR takes other predictors into account. The interpretation for an estimated coefficient in MLR is the effect on the dependent variable after changing the given predictor by 1 unit *holding all other variables constant*, while the interpretation of an estimated coefficient in SLR is simply the effect on the dependent variable after changing the given predictor by 1 unit.

c) Yes, adding the additional predictor Girth into the model is helpful in explaining the variability in Volume, as the R2 term when regressing Height onto Volume is only 0.3579, while the R2 term when including both Height and Girth in the model for Volume is 0.948. Also, the correlation between Girth and Height is not particularly strong (r = 0.519), so there isn’t a tremendous amount of redundant information in Girth and Height when predicting Volume.

#4. a) > anova(fit.slr, fit.mlr)

Analysis of Variance Table

Model 1: Volume ~ Height

Model 2: Volume ~ Height + Girth

Res.Df RSS Df Sum of Sq F Pr(>F)

1 29 5204.9

2 28 421.9 1 4783 317.41 < 2.2e-16 \*\*\*

b) 1) H0: BGirth = 0

Ha: BGirth ≠ 0

2) F = 317.41

Numerator df = q = 1; Denominator df = n – k – 1 = 28

p-value < 2.2 x 10-16

3) We reject that BGirth = 0, meaning that the partial effect of Girth is significant in explaining Volume given Height is already in the model (i.e. use the full model).

c) Yes, the partial F-test for Girth is consistent with the t-test for the Girth coefficient in the MLR model, as an F distribution with 1 numerator df and *v* denominator df is equal to a t distribution with *v* df squared. This result is consistent with these particular tests (since q, the subset of predictors we are taking out, is 1 in this case), as (17.816)2 = 317.41.

#5. a) > fit.V.H = lm(Volume ~ Height)

> fit.G.H = lm(Girth ~ Height)

> fit.part.reg = lm(fit.V.H$residuals ~ fit.G.H$residuals)

> summary(fit.part.reg)

Coefficients:

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

(Intercept) -4.806e-15 6.851e-01 0.00 1

fit.G.H$residuals 4.708e+00 2.597e-01 18.13 <2e-16 \*\*\*

Residual standard error: 3.814 on 29 degrees of freedom

Multiple R-squared: 0.9189, Adjusted R-squared: 0.9161

F-statistic: 328.7 on 1 and 29 DF, p-value: < 2.2e-16

> plot(fit.G.H$residuals, fit.V.H$residuals, pch=16)

> abline(fit.part.reg)



\* Girth can explain 91.89% of the variability left unexplained by Height, which can be seen by the R2 in the R summary output for the SLR of the Girth vs Height residuals onto the Volume vs Height residuals.

b) R2MLR = R2SLR + R2PLR(1 – R2SLR)

= 0.3579 + 0.9189(1 – 0.3579)

= 0.3579 + 0.9189(0.6421)

= 0.3579 + 0.59

R2MLR = 0.948

c) The estimated slope in the partial regression plot corresponds to Girth’s regression coefficient estimate in the MLR.