**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 4 Solutions**

Prepare a short report with relevant output, your comments, and answers to the questions (this does not need to be exhaustive or polished, but should contain enough to show that you completed all tasks and analyses).

Submit the report at the end of the lab session.

Consider the dataset *bears.txt*.

This contains several variables measured on n=141 “bear capturing” occasions, with the following variables:

*ID:* Identification number

*Age:* Bear's age, in months

*Month:* Month when the measurement was made. Sex. 1 = male 2 = female

*Head.L:* Length of the head, in inches

*Head.W:* Width of the head, in inches

*Neck.G:* Girth (distance around) the neck, in inches

*Length:* Body length, in inches

*Chest.G:* Girth (distance around) the chest, in inches

*Weight:* Weight of the bear, in pounds

*Obs.No:* Observation number for this bear. For example, the bear with ID=41 (Bertha) was measured on four occasions. The value of Obs.No goes from 1 to 4 for these observations

*Name:* The names of the bears given to them by the researchers.

The observations are not independent, because the same bear may have been captured more than once (see variables “Name”, “ID” and “Obs.No”).

* For each bear, select only the first observation, so that the new dataset will contain only independent observations. Why is that important for linear regression? How many bears do we have in the dataset?

> bears=read.csv("bears.txt", header=T, sep="")

> bears=bears[bears$Obs.No==1,]

> head(bears)

ID Age Month Sex Head.L Head.W Neck.G Length Chest.G Weight Obs.No Name

1 598 NA 4 1 13.5 7.0 24.5 62.0 41 248 1 Albert

2 578 NA 4 1 18.5 8.5 23.5 67.5 42 204 1 Bill

5 179 100 4 2 13.0 7.0 21.0 70.0 41 220 1 Fannie

7 253 51 4 1 13.5 8.0 27.0 68.5 49 360 1 John

8 47 NA 4 1 15.5 7.0 29.3 76.0 53 416 1 Palmer

9 592 NA 4 2 13.0 7.0 21.0 59.0 34 146 1 Vanessa

**Independence between observations is important for linear regression because if observations are not independent, which means these observations have been counted for more than one times, then these observations will be weighted more than others, which will influence our accuracy of the linear regression model.**

**There are 99 bears in this data set.**

Consider the variables y=“Weight”, x1=“Chest.G” and x2=“Head.W”.

Fit two separate simple regression models for y=“Weight” on x1=“Chest.G”, and y=“Weight” on x2=“Head.W” (you can use the *lm* function or the equations).

* Do the estimated regression slopes suggest positive or negative relationships? Is there a meaningful interpretation for the regression intercepts?

**Model 1:**

> model1=lm(Weight~Chest.G)

> summary(model1)

Call:

lm(formula = Weight ~ Chest.G)

Residuals:

Min 1Q Median 3Q Max

-77.75 -18.32 -0.63 17.22 97.78

Coefficients:

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

(Intercept) -266.6770 13.2722 -20.09 <2e-16 \*\*\*

Chest.G 12.6462 0.3586 35.27 <2e-16 \*\*\*

---

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

Residual standard error: 30.77 on 97 degrees of freedom

Multiple R-squared: 0.9276, Adjusted R-squared: 0.9269

F-statistic: 1244 on 1 and 97 DF, p-value: < 2.2e-16

**Model 2:**

> model2=lm(Weight~Head.W)

> summary(model2)

Call:

lm(formula = Weight ~ Head.W)

Residuals:

Min 1Q Median 3Q Max

-186.60 -40.84 -11.71 26.70 223.84

Coefficients:

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

(Intercept) -201.697 34.906 -5.778 9.13e-08 \*\*\*

Head.W 61.482 5.372 11.446 < 2e-16 \*\*\*

---

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

Residual standard error: 74.61 on 97 degrees of freedom

Multiple R-squared: 0.5746, Adjusted R-squared: 0.5702

F-statistic: 131 on 1 and 97 DF, p-value: < 2.2e-16

**Both the estimated regression slopes suggest positive relationships.**

**Since these intercepts have nothing to do to interpret the correlation between variables, there is no meaningful interpretation for these intercepts.**

* Using the equation, estimate the variance sigma2 of the error term for the two models (you can check the result with the *summary* function, but you need to compute it using the equation).

> sigma2.1=sum((model1$residuals)^2)/97

> sigma2.2=sum((model2$residuals)^2)/97

> sigma2.1

[1] 946.6225

> sigma2.2

[1] 5566.195

**The variance σ^2 of model 1 is 946.6225.**

**The variance σ^2 of model 2 is 5566.195.**

* Using the equation, compute the coefficient of determination R2 for both regressions (you can check the result with the *summary* function, but you need to compute it using the equation). What is their interpretation?

> RSS.1=sum(model1$residuals^2)

> RSS.2=sum(model2$residuals^2)

> TotalSS=sum((bears$Weight - mean(bears$Weight))^2)

> R2.1=1-RSS.1/TotalSS

> R2.2=1-RSS.2/TotalSS

> R2.1

[1] 0.9276481

> R2.2

[1] 0.5745668

**The R^2 of model 1 is 0.9276481, which shows a really strong correlation between variables Weight and Chest.G because the value of R^2 is close to 1, which means a perfect linear relation.**

**The R^2 of model 2 is 0.5745668, which shows that variables Weight and Head.W have linear relationship but not as strong as model 1.**

* Between x1=“Chest.G” and x2=“Head.W”, which appears to be the best predictor for y=“Weight”? (Address this comparing the coefficients of determination R2 of the two regressions).

|  |  |
| --- | --- |
| |  | | --- | |  | |
| |  | | --- | |  | |

**Between x1 and x2, x1 appears to be the best predictor for y because the coefficients of determination R^2 of x1 is 0.9276481, which is much larger than that of x2.**

The following part will NOT be considered in grading the lab report.

Fit a multiple linear regression model with predictors x1=“Chest.G” and x2=“Head.W”.

* Using the equation, estimate the variance sigma2 of the error term for the new model (you can check the result with the *summary* function, but you need to compute it using the equation).
* Using the equation, compute the coefficient of determination R2 for the new regression (you can check the result with the *summary* function, but you need to compute it using the equation). What is its interpretation?
* Do you think this model is better that the one with only x1? Why?

R code:

setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/math462")

getwd()

bears=read.csv("bears.txt", header=T, sep="")

bears=bears[bears$Obs.No==1,]

head(bears)

attach(bears)

#model1

model1=lm(Weight~Chest.G)

summary(model1)

#model2

model2=lm(Weight~Head.W)

summary(model2)

sigma2.1=sum((model1$residuals)^2)/97

sigma2.2=sum((model2$residuals)^2)/97

sigma2.1

sigma2.2

RSS.1=sum(model1$residuals^2)

RSS.2=sum(model2$residuals^2)

TotalSS=sum((bears$Weight - mean(bears$Weight))^2)

R2.1=1-RSS.1/TotalSS

R2.2=1-RSS.2/TotalSS

R2.1

R2.2