**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 5**

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 again the dataset *bears.txt* used in Lab4.

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.

As you did in Lab4, consider only the first observation for each bear (bears\_indep=bears[bears$Obs.No==1,]).

Consider the simple linear regression model with response y=“Weight” and predictor x1=“Neck.G” (you can fit the model using the *lm* function or the equations).

* Provide the estimated regression slope. Does it suggest positive or negative relationship?

> model=lm(Weight~Neck.G)

> summary(model)

Call:

lm(formula = Weight ~ Neck.G)

Residuals:

Min 1Q Median 3Q Max

-97.011 -19.446 -3.831 15.644 168.594

Coefficients:

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

(Intercept) -244.7885 15.4956 -15.8 <2e-16 \*\*\*

Neck.G 20.5900 0.7148 28.8 <2e-16 \*\*\*

---

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

Residual standard error: 37.01 on 97 degrees of freedom

Multiple R-squared: 0.8953, Adjusted R-squared: 0.8942

F-statistic: 829.7 on 1 and 97 DF, p-value: < 2.2e-16

Regression slope=20.59

It suggests positive relationship.

* Using the equation, estimate the variance of the errors (i.e. compute sigma2\_hat) and the standard error of the estimated slope (i.e. compute se(beta1\_hat) ). You can check the result with the *summary* function, but you need to compute it using the equation.

> y=Weight

> X=model.matrix(Weight~Neck.G, data=bears)

> n=nrow(X)

> p=ncol(X)

> df=n-p

> df

[1] 97

> beta.hat=solve(t(X)%\*%X,t(X)%\*%y)

> beta.hat

[,1]

(Intercept) -244.78848

Neck.G 20.58996

>

> y.hat=X%\*%beta.hat

> res=y-y.hat

> sigma2.hat=sum(res^2)/(n-p)

> sigma2.hat

[1] 1369.55

>

> cov.mat=sigma2.hat\*solve(t(X)%\*%X)

> round(cov.mat,digits=4)

(Intercept) Neck.G

(Intercept) 240.1124 -10.7529

Neck.G -10.7529 0.5110

> se=sqrt(diag(cov.mat))

> mytable=cbind(beta.hat,se)

> colnames(mytable)<-c("estimate","se")

> mytable

estimate se

(Intercept) -244.78848 15.4955592

Neck.G 20.58996 0.7148337

The variance of the errors is 1369.55, standard error of the estimated slope is 0.7148337.

* Perform a test for the hypotheses: H0: beta1 = 0 vs H1: beta1 != 0. Provide:
* Distribution of the test statistic under H0 (type and parameters of the distribution)
* Value of the test statistics, computed using the equation (you can check the result with the *summary* function, but you need to compute it using the equation)
* P-value, computed using the equation (you can check the result with the *summary* function, but you need to compute it using the equation)
* Interpretation of the result.

Fit a multiple linear regression model with response y=“Weight” and predictors x1=“ Neck.G” and x2=“Head.W” (you can fit the model using the *lm* function or the equations).

**I am using the t distribution with t(97).**

> t=beta.hat/se

> pvalue=2\*pt(abs(t),df=n-p,lower.tail=FALSE)

> mytable=cbind(mytable,t,pvalue)

> colnames(mytable)<-c("estimate","se","t-statistics","p-value")

> mytable

estimate se t-statistics p-value

(Intercept) -244.78848 15.4955592 -15.79733 1.433795e-28

Neck.G 20.58996 0.7148337 28.80384 2.475672e-49

By using the equation, we have t=28.80384, p-value=2.475672e-49

Since p-value=2.475672e-49 < α, we reject H0.

We can conclude that beta is not equal to 0, which means the predictor has correlation with the response.

* Using the equation, compute TSS, RSS and SS\_reg.

> TSS=sum((y-mean(y))^2)

> TSS

[1] 1269109

>

> y.hat.SS=bears.SS$fitted.values

> RSS=sum((y-y.hat.SS)^2)

> RSS

[1] 132756.1

>

> SS.reg=TSS-RSS

> SS.reg

[1] 1136353

**TSS=1269109**

**RSS=132756.1**

**SS\_reg=1136353**

* Perform a test on all the predictors, i.e. test the hypotheses H0: beta1 = beta2 = 0 vs H1: at least one != 0. Provide:
* Distribution of the test statistic under H0 (type and parameters of the distribution)
* Value of the test statistics, computed using the equation (you can check the result with the *summary* function, but you need to compute it using the equation)
* P-value, computed using the equation (you can check the result with the *summary* function, but you need to compute it using the equation)
* Interpretation of the result.

**I am using the F distribution with F(2,96).**

> n.reg=length(y)

> p.reg=length(bears.SS$coefficients)

> n.reg

[1] 99

> p.reg

[1] 3

> F=(SS.reg/(p.reg-1))/(RSS/(n.reg-p.reg))

> F

[1] 410.8658

>

> pvalue=pf(F,p.reg-1,n.reg-p.reg,lower.tail=FALSE)

> pvalue

[1] 8.682892e-48

F=410.8658

pvalue=8.682892e-48

By using the equation, we have F=410.8658, p-value=8.682892e-48

Since pvalue=8.682892e-48 < α, we reject H0.

We can conclude that at least one beta is not equal to 0, which means at least one of the predictor has correlation with the response.