APPLIED STATISTICS
TUTORIAL 5 SOLUTIONS

Question 3 in Tutorial 4 (Con'd, revised based on ex 10.09 from "The Statistical Sleuth")

As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species. This data is contained in the file "crab.csv".

a) Fit a regression model of log(force) on log(height) and species, allow for an interaction between log(height) and species. Let Hemigrapsus nududus be the baseline species, i.e., do not use an indicator variable for this species.

```
>crab<-read.table("crab.csv",header=T,sep=",")
>names(crab)
>force=crab$FORCE
>height=crab$HEIGHT
>species=crab$SPECIES
>ILP=ifelse(species==species[16],1,0)
>ICP=ifelse(species==species[28],1,0)
>crab.reg=lm(log(force)~log(height)+ILP+ICP+ILP*log(height)+ICP*log(height))

Call: lm(formula = log(force) ~ log(height) + ILP + ICP + ILP * log(height) + ICP *
   log(height))
Residuals:
     Min      1Q   Median     3Q    Max
 -0.7668 -0.2851 -0.02306 0.2425 0.8882

Coefficients:
                  Value Std. Error t value Pr(>|t|)
    (Intercept)  0.5191  1.0001     0.5191  0.6073
    log(height)  0.4083  0.4868     0.8387  0.4079
            ILP -4.2992  1.5283    -2.8131  0.0083
            ICP -2.4864  1.7606    -1.4123  0.1675
ILP:log(height)  2.5653  0.7354     3.4885  0.0014
ICP:log(height)  1.6601  0.7889     2.1043  0.0433

Residual standard error: 0.4329 on 32 degrees of freedom
Multiple R-Squared: 0.7945
F-statistic: 24.75 on 5 and 32 degrees of freedom, the p-value is 3.935e-010

Correlation of Coefficients:
                (Intercept) log(height)     ILP     ICP ILP:log(height)
    log(height) -0.9933
            ILP -0.6544      0.6500
            ICP -0.5680      0.5642      0.3717
ILP:log(height)  0.6576     -0.6620     -0.9937 -0.3735
ICP:log(height)  0.6130     -0.6171     -0.4011 -0.9934  0.4085
```

b) What is the p-value for the test of the hypothesis that the slope in the regression of log(force) on log(height) is the same for Lophopanopeus bellus as it is for Hemigrapsus nududus?

We need to test whether $\beta_4=0$. β_4 gives the difference in slope for the species Lophopanopeus bellus and Hemigrapsus nududus. From the output in (a) we can see that the two-sided p-value is 0.0014 (reject null that $\beta_4=0$). The data suggests that the slopes are different.

c) What is a 95% CI for the amount by which the slope for Cancer productus exceeds the slope for Hemigrapsus nududus?

Now we are interested in β_5. β_5 gives the difference in slope for the species Cancer productus and Hemigrapsus nududus. The estimate of β_5 is 1.66 and the SE of this estimate is 0.79. For a 95% CI we need to find t(32,0.975)=2.037 [using 2 is here is okay]

$$CI=(1.66-2.037*0.79,1.66+2.037*0.79)=(0.053,3.267)$$

This gives a plausible range for β_5. This range suggests that β_5 is different from 0.

d) Is the regression model fit in (a) significant? Provide a p-value for the test.

The F-test for the overall significance of the regression is given in the output from (a).

```
F-statistic: 24.75 on 5 and 32 degrees of freedom, the p-value is 3.935e-010
```

The extremely small p-value means the regression is highly significant.

e) Are the slopes of the regression lines the same for the three species? (Hint: You will need to use the anova() command and an F-test)

To answer this question we need to test whether $\beta_4=\beta_5=0$.

```
> crab.regr=lm(log(force)~log(height)+ILP+ICP) #reduced model
> anova(crab.regr,crab.reg,test='F')
Analysis of Variance Table

Model 1: log(force) ~ log(height) + ILP + ICP
Model 2: log(force) ~ log(height) + ILP + ICP + ILP * log(height) + ICP *
    log(height)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     34 8.3816
2     32 5.9971  2    2.3844 6.3615 0.00472 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null ($\beta_4=\beta_5=0$) and conclude that the slopes are different.

Question 1 (revised based on the exercise in Chapter 10 from "The Statistical Sleuth")

The Old Faithful data used in class also contains a column called DATE. This column contains information on the day the data were collected. The data is contained in "oldfaithful.csv". Fit the regression of interval on duration and date (use seven indicator variables to distinguish the eight dates). Construct an F-statistic for the test of whether any difference in mean intervals is due to the date of recording. Find the p-value.

```
>old=rab<-read.table("oldfaithful.csv",header=T,sep=",")
>names(old)
>date=old$DATE
>duration=old$DURATION
>interval=old$INTERVAL
>I2=ifelse(date==2,1,0)
>I3=ifelse(date==3,1,0)
>I4=ifelse(date==4,1,0)
>I5=ifelse(date==5,1,0)
>I6=ifelse(date==6,1,0)
>I7=ifelse(date==7,1,0)
```

```
>I8=ifelse(date==8,1,0)
>old.reg=lm(interval~duration+I2+I3+I4+I5+I6+I7+I8)
> old.regr=lm(interval~duration) #reduced model
> anova(old.regr,old.reg,test='F')
Analysis of Variance Table

Model 1: interval ~ duration
Model 2: interval ~ duration + I2 + I3 + I4 + I5 + I6 + I7 + I8
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    105 4689.0
2     98 4620.2  7    68.853 0.2086 0.9828
```

We cannot reject the null ($\beta_2= \beta_3=\ldots= \beta_8=0$).