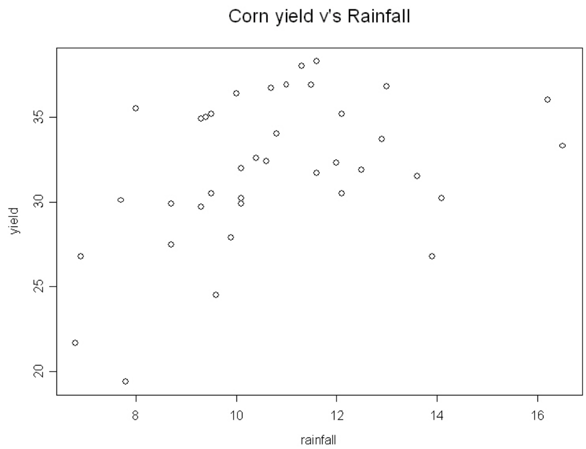
APPLIED STATISTICS
TUTORIAL 4 SOLUTIONS

Question 1 (revised based on ex 9.15 from "The Statistical Sleuth")

a) Plot corn yield versus rainfall.

```
>corn<-read.table("corn.csv",header=T,sep=",")
>names(corn)
>year=corn$YEAR
>yield=corn$YIELD
>rain=corn$RAIN
>plot(rain,yield,main="Corn yield v's Rainfall",ylab="yield",xlab="rainfall")
```



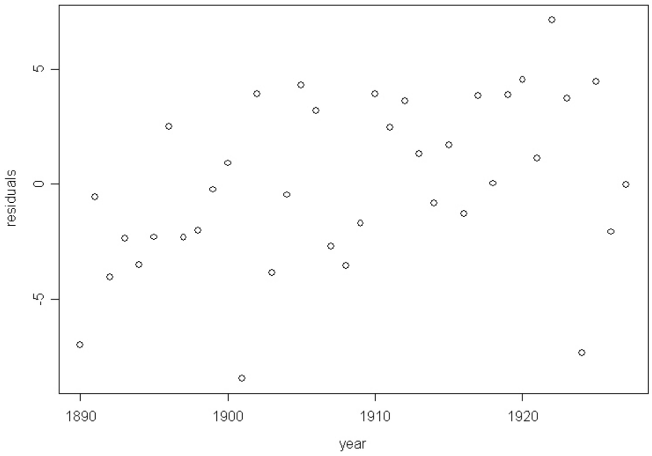
b) Fit the multiple linear regression of corn yield on rain and rain^2.

```
>corn.reg=lm(yield~rain+I(rain^2))
>summary(corn.reg)

Call: lm(formula = yield ~ rain + rain^2)
Residuals:
    Min     1Q  Median    3Q  Max
 -8.464 -2.324 -0.1265 3.515 7.16

Coefficients:
               Value  Std. Error  t value  Pr(>|t|)
(Intercept)  -5.0147  11.4416    -0.4383   0.6639
       rain   6.0043   2.0389     2.9448   0.0057
  I(rain^2)  -0.2294   0.0886    -2.5877   0.0140

Residual standard error: 3.763 on 35 degrees of freedom
Multiple R-Squared: 0.2967
F-statistic: 7.382 on 2 and 35 degrees of freedom, the p-value is 0.002115

Correlation of Coefficients:
          (Intercept)    rain
     rain -0.9906
I(rain^2)  0.9648     -0.9910
```

The fitted regression line is: $\hat{\mu}$ (yield|rain,rain^2) = -5+6.0rain-0.23rain^2.

c) Plot the residuals versus year. Is there pattern evident in this plot? What does it mean

```
>plot(year,corn.reg$residuals,ylab="residuals",xlab="year")
```



There is a trend in the residuals. The residuals tend to increase as year increases. This trend suggests that we investigate including year in the fitted model.

d) Fit the multiple regression of corn yield on rain, rain^2, and year. How do the coefficients of rain and rain^2 differ from those in the estimated model in (b)? How does the estimate of σ differ? How do the standard errors of the coefficients differ? Describe the effect of an increase of one inch of rainfall on the mean yield over the range of rainfalls and years.

```
>cornyear.reg = lm(yield ~ rain + I(rain^2) + year)
>summary(cornyear.reg)

Call: lm(formula = yield ~ rain + I(rain^2) + year)
Residuals:
  Min      1Q   Median     3Q   Max
 -9.4 -1.809 -0.04788 2.405 5.184

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -263.3032   98.2410    -2.6802    0.0113
       rain    5.6704    1.8882     3.0030    0.0050
  I(rain^2)   -0.2155    0.0821    -2.6259    0.0129
       year    0.1363    0.0516     2.6445    0.0123

Residual standard error: 3.477 on 34 degrees of freedom
Multiple R-Squared: 0.4167
F-statistic: 8.095 on 3 and 34 degrees of freedom, the p-value is 0.0003339

Correlation of Coefficients:
          (Intercept)    rain I(rain^2)
     rain -0.0399
I(rain^2)  0.0401     -0.9910
     year -0.9942     -0.0669  0.0639
```

The fitted regression is: $\hat{\mu}$(yield|rain,rain^2,year) = -263+5.7rain-0.22rain^2+0.14year.

The estimates are similar but the standard errors are a slightly smaller. The estimates between the two models do not change much. The reason for this is that year and rainfall are not very highly correlated. The standard errors are smaller because the additional variable (year) is an important variable.

e) Fit the multiple regression of corn yield on rain, rain^2, year, and year$\times$rain. Is the coefficient of the interaction term significantly different from zero? Interpret the interaction term?

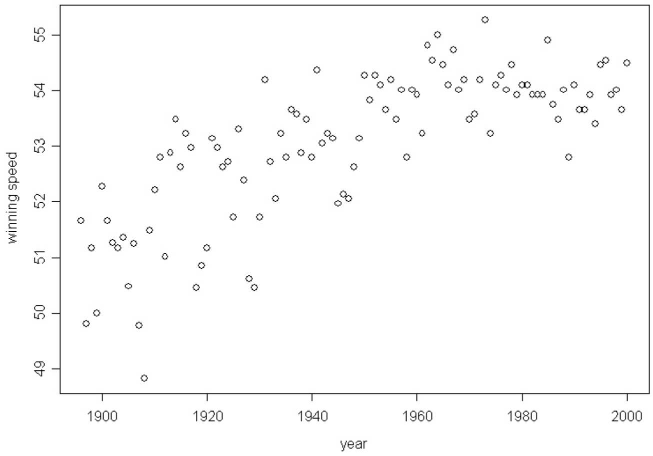
```
>cornint.reg=lm(yield~rain+I(rain^2)+year+I(rain*year))
>summary(cornint.reg)

Call: lm(formula = yield ~ rain + I(rain^2) + year + I(rain * year))
Residuals:
    Min      1Q Median     3Q  Max
 -6.297 -2.547 0.6011 1.992 5.02

Coefficients:
                 Value  Std. Error    t value   Pr(>|t|)
(Intercept) -1909.4647   486.2435    -3.9270     0.0004
       rain   158.8411    44.5681     3.5640     0.0011
  I(rain^2)    -0.1862     0.0720    -2.5876     0.0143
       year     1.0012     0.2554     3.9193     0.0004
  rain:year    -0.0806     0.0234    -3.4391     0.0016

Residual standard error: 3.028 on 33 degrees of freedom
Multiple R-Squared: 0.5706
F-statistic: 10.96 on 4 and 33 degrees of freedom, the p-value is 9.127e-006

Correlation of Coefficients:
          (Intercept)     rain I(rain^2)     year
     rain -0.9840
I(rain^2) -0.1093      0.0818
     year -0.9998      0.9833  0.1275
rain:year  0.9844     -0.9993 -0.1182   -0.9844
```

The two sided p-value for the significance of the interaction term is 0.0016. Because of the negative coefficient value (-0.0806), this indicates that the effect of rainfall on yield is smaller for years closer to 1927.

Question 2 (revised based on ex 9.20 from "The Statistical Sleuth")

The Kentucky Derby is a 1.25 mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky. The file "derby.csv" contains the data on the year of the race, the winning horse, the conditions of the track, and the average speed (in feet per second) of the winner, for years 1896-2000. The track conditions have been grouped into three categories: fast, good, and, slow. Model the mean winning speed as a function of year and track conditions. (Hint: The first thing you should do is look at a plot of winning speed versus year. What does a curved plot suggest?)

```
>derby<-read.table("derby.csv",header=T,sep=",")
>names(derby)
>year=derby$year
>speed=derby$speed
>condition=derby$condition
>plot(year,speed,xlab="year",ylab="winning speed")
```



The curved nature of this plot suggests we should include $year^2$. Using "fast" as the baseline track condition we fit the following MLR:

$$\mu(speed|year,condition)=\beta_0+\beta_1 year+\beta_2 year^2+\beta_3 I(slow)+\beta_4 I(good)$$

I(slow) is an indicator variable taking the value "1" when track condition is slow and "0" otherwise, likewise for I(good).

```
>Igood=ifelse(condition==condition[1],1,0)
>Islow=ifelse(condition==condition[2],1,0)
>derby.reg=lm(speed~year+I(year^2)+Igood+Islow)
>summary(derby.reg)

Call: lm(formula = speed ~ year + year^2 + Igood + Islow)
Residuals:
    Min     1Q   Median     3Q Max
 -1.609 -0.308 -0.02224 0.3885 1.1

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1597.6374   247.6026    -6.4524     0.0000
       year     1.6686     0.2543     6.5626     0.0000
  I(year^2)    -0.0004     0.0001    -6.4569     0.0000
      Igood    -0.5319     0.1862    -2.8574     0.0052
      Islow    -1.6099     0.1439   -11.1890     0.0000

Residual standard error: 0.5492 on 100 degrees of freedom
Multiple R-Squared: 0.8365
F-statistic: 127.9 on 4 and 100 degrees of freedom, the p-value is 0

Correlation of Coefficients:
          (Intercept)    year I(year^2)   Igood
     year -1.0000
I(year^2)  0.9999     -1.0000
    Igood  0.0240     -0.0249  0.0257
    Islow -0.0098      0.0076 -0.0056    0.1817
```

The fitted model is:

$$\hat{\mu}(speed|year,condition)=1597+1.67year+-0.0004year^2-1.61I(slow)-0.53I(good)$$

Question 3 (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.