APPLIED STATISTICS
SOLUTIONS: TUTORIAL 2

Question 1 (revised based on ex 7.27 from Statistical Sleuth)
The file "ex0727.csv" contains measured distances and recession velocities for 10 clusters of nebulae. According to a theory by Hubble the mean of the measured distance, as a function of velocity, should be β_1*velocity (i.e., μ(distance|velocity) = β_1*velocity), and β_1 is the age of the universe.

b) Using Hubble's theory what is the estimated age of the universe? (Hint: The function lm() includes an intercept by default. lm(Y~X-1) fits the SLR without an intercept, i.e., $\mu(Y|X)= \beta_1 X$.

To use the hubble theory we need to fit the above regression without the intercept term.

```
nointhub.reg=lm(distance~velocity-1)  #fitting SLR with no intercept.
summary(nointhub.reg)
summary(nointhub.reg)

Call:
lm(formula = distance ~ velocity - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9045  0.1663  0.6608  1.0017  1.9530

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
velocity 1.730e-03  4.769e-05   36.27 4.56e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.9932,     Adjusted R-squared: 0.9925
F-statistic:  1316 on 1 and 9 DF,  p-value: 4.558e-11
```

The fitted regression line is $\hat{\mu}$(distance|velocity) = 0.0017*velocity.

Using the β_1 estimate of 0.0017, the estimated aged of the universe is 1.70 billion years.

Question 2 (revised based on ex 7.29 from Statistical Sleuth)
Black wheatears are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. A study was conducted (M. Soler et al.) which calculated the average stone mass (g) carried by each of 21 male wheatears, along with T-cell response measurements reflecting their immune systems' strengths. The file "ex0729.csv" contains the data.

c) For wheatears that carry stones with an average mass of 2g, what would you estimate their mean T-cell response to be? Comment on this estimate.

0.0875+0.0328*2

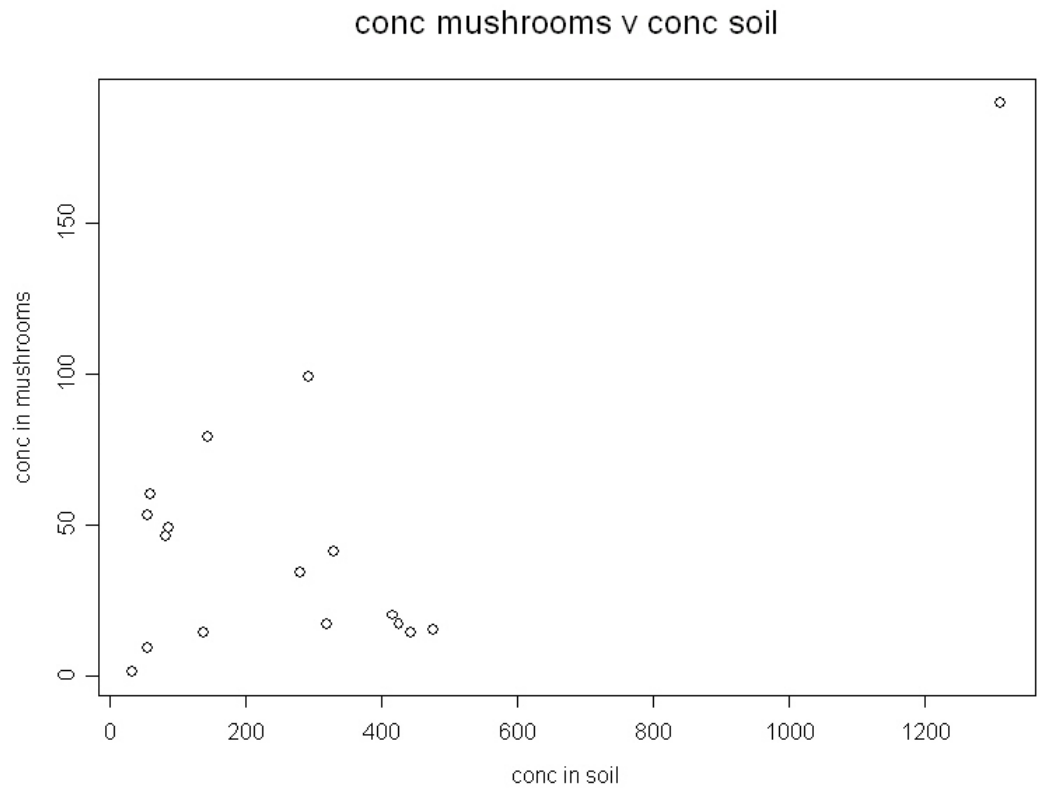
The minimum explanatory variable value in our sample is 3.33. It is dangerous to make statements outside of the range of our explanatory variable (extrapolation). The straight line model is not necessarily valid over a wider range of explanatory variable values.

Question 3 (revised based on ex 8.18 from Statistical Sleuth)

One of the most dangerous contaminants deposited over European countries following the Chernobyl accident of April 1987 was radioactive cesium. To study cesium transfer from contaminated soil to plants, researchers collected soil samples and samples of mushroom mycelia from 17 wooded locations in Umbria, Central Italy from August 1986 to November 1989. Measured concentrations of cesium (Bq/Kg) in the soil and in the mushrooms are contained in the file "ex0818.csv".

a) Construct a scatterplot of Y = concentration in mushrooms and X = concentration in soil. What do you notice?

```
chernobyl<-read.table("ex0818.csv",header=T,sep=",")
names(chernobyl)
mush=chernobyl$MUSHROOM
soil=chernobyl$SOIL
plot(soil,mush, ylab="conc in mushrooms", xlab="conc in soil", main="conc mushrooms
    conc soil")
```



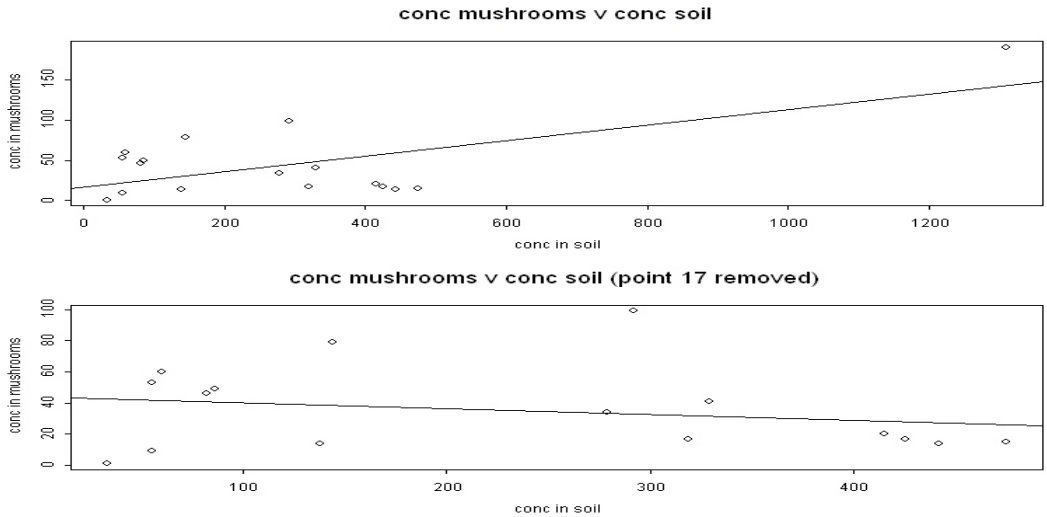
There is one point that is far removed from the rest of the data. This point is point number 17.

b) Fit a simple linear regression using Y = concentration in mushrooms and X = concentration in soil. Produce a plot of the fitted regression line superimposed on the scatterplot of points.

c) Repeat part (b) with point 17 removed. What do you notice?

Answer to part (b) and (c) combined.

```
par(mfrow=c(2,1))
plot(soil,mush, ylab="conc in mushrooms", xlab="conc in soil", main="conc mushrooms v
    conc soil")
abline(lm(mush~soil))
plot(soil[-17],mush[-17], ylab="conc in mushrooms", xlab="conc in soil", main="conc
    mushrooms v conc soil (point 17 removed)")
abline(lm(mush[-17]~soil[-17]))
```



The regression is highly sensitive to point 17. The fitted regression line changes dramatically according to whether point 17 is excluded or included. Point 17 is said to be highly influential. In this example it is probably better to superimpose the regression line with point 17 removed on the original plot. This can be done by typing the command *abline(lm(mush[-17]~soil[-17]))* straight after the command *abline(lm(mush~soil))*.

Question 4

Part 2. PROOF.

We first show Part 2. Since $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, we have

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y},$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

The second equation is trivial based on this result.

Part 1. PROOF.

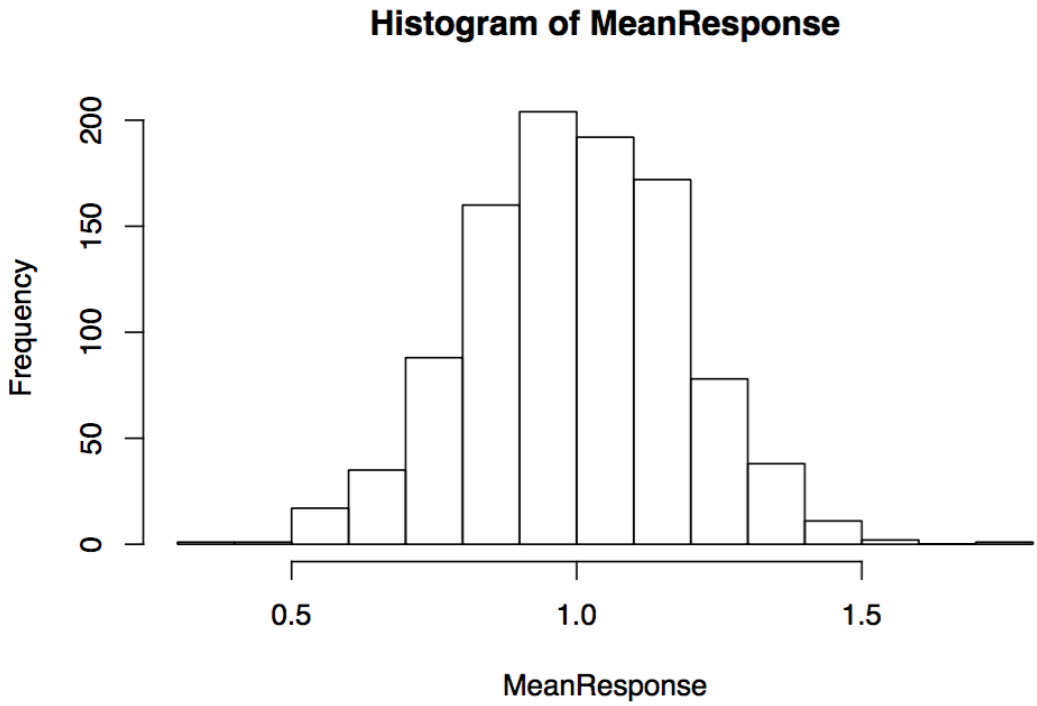
$$\overline{\text{res}} = \frac{1}{n}\sum_{i=1}^{n} \text{res}_i = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \bar{Y} - \bar{\hat{Y}} = 0,$$

due to the result of Part 2.

The second equation is trivial based on this result.

Part 3.

```
X=1:100
n=length(X)
Y=rep(0,n)
numsamp=1000
MeanResponse=rep(0,numsamp)
Xnew=2.5
set.seed(1)
for(i in 1:numsamp) {
  errors=rnorm(n)
  Y=1+0*X+errors
  SLRfit=lm(Y~X)
  MeanResponse[i]=SLRfit$coef[1]+SLRfit$coef[2]*Xnew
}
hist(MeanResponse)
```




```
mean(MeanResponse)
```

```
[1] 0.9979859
```

```
sd(MeanResponse)
```

```
[1] 0.1877534
```