Solutions Tutorial 3

Question 1 in Tutorial 2 (Con'd, 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 $\beta_1 *$velocity (i.e., μ(distance|velocity) = $\beta_1 *$velocity), and β_1 is the age of the universe.

a) Are the data consistent with the theory that the intercept (β_0) is zero?

```
hubble<-read.table("ex0727.csv",header=T,sep=",")
distance=hubble$DISTANCE
velocity=hubble$VELOCITY
hub.reg=lm(distance~velocity)   #fitting the SLR with an intercept term
summary(hub.reg)

Call:
lm(formula = distance ~ velocity)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.57390 -0.61564  0.08178  0.43318  1.65894 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.537e-01  5.251e-01   1.816    0.107    
velocity    1.643e-03  6.401e-05  25.668 5.69e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.104 on 8 degrees of freedom
Multiple R-squared: 0.988,      Adjusted R-squared: 0.9865 
F-statistic: 658.9 on 1 and 8 DF,  p-value: 5.691e-09
```

The fitted regression line is $\hat{\mu}$(distance|velocity) = 0.9537+0.0016$*$velocity. To see whether the data is consistent with the theory we need to test null: $\beta_0 = 0$ v's alt: $\beta_0 \neq 0$. From the summary output it is given that the test statistic for this hypothesis is 1.82 with a p-value of 0.1069. There is no reason to reject the theory (we cannot reject the null).

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.

c) Produce a 95% confidence interval for the estimate in part (b).

Our estimate of β_1 was 0.0017 with a SE of 0.000048. A 95% CI for β_1 is given by:

(0.0017-t(0.975,9)*0.000048, 0.0017+t(0.975,9)*0.000048)
=(0.0016, 0.0018)
converting this CI to years gives (1.59 billion, 1.80 billion)

Note:
(1) the SE was hard to see in the original summary output (it was too small!). The following command shows us the SE:

```
summary(nointhub.reg)$coefficients

               Estimate   Std. Error t value     Pr(>|t|)
velocity 0.001729822 4.769186e-05 36.2708 4.557676e-11
```

(2) t(0.975,8) was found using the command: `qt(0.975,length(distance)-1)`

Question 2 in Tutorial 2 (Con'd, 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.

a) Does the data suggest there is a linear relationship between the mean T-cell response and average stone mass?

```
wheatears<-read.table("ex0729.csv",header=T,sep=",")
tcell=wheatears$tcell
mass=wheatears$mass
wheatears.reg=lm(tcell~mass)
summary(wheatears.reg)

Call:
lm(formula = tcell ~ mass)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18138 -0.04673  0.01796  0.04219  0.15999 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.08750    0.07868   1.112  0.27996   
mass         0.03282    0.01064   3.084  0.00611 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.08102 on 19 degrees of freedom
Multiple R-squared: 0.3336,     Adjusted R-squared: 0.2986 
F-statistic: 9.513 on 1 and 19 DF,  p-value: 0.006105
```

We want to test null: $\beta_1 = 0$ v's alt: $\beta_1 \neq 0$. From the summary output we are given that the p-value for this test is 0.0061. We reject the null and conclude that there is a linear relationship.

b) If a particular wheatear carries stones that have an average mass of 6.5g, what would you predict its T-cell response to be? Provide a 90% prediction interval for this prediction.

The fitted regression line is $\hat{\mu}$(T-cell|mass) = 0.0875+0.0328*mass. This means we would predict its T-cell response to be 0.0875+0.0328*6.5=0.30. Recall the PI formula is:

$$(b_0+b_1X_0) \pm t(n-2,100-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)s_x^2}}$$

We can obtain all the pieces from R:
From the summary() output above we have: n=21, $\hat{\sigma}$=0.081.
The other pieces are calculated below:

```
> mean(mass)
[1] 7.204286
> sqrt(var(mass))
[1] 1.702441
> qt(0.95, length(mass) - 2)
[1] 1.729133
```

The 90% PI is:

$$(0.0875+0.0328*6.5) \pm 1.73*0.081\sqrt{1+\frac{1}{21}+\frac{(6.5-7.2)^2}{(21-1)(1.7)^2}}$$

Note: Explore the R function "predict" for an automated way to produce both confidence and prediction intervals (the R codes in Lecture Notes 2).

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 1 (revised based on ex 8.23 from Statistical Sleuth)
The data in the file "ex0823.csv" are the average wine consumption rates (in litres per person) and number of ischemic heart disease deaths (per 1000 men aged 55 to 64 years old) for 18 industrialised countries.

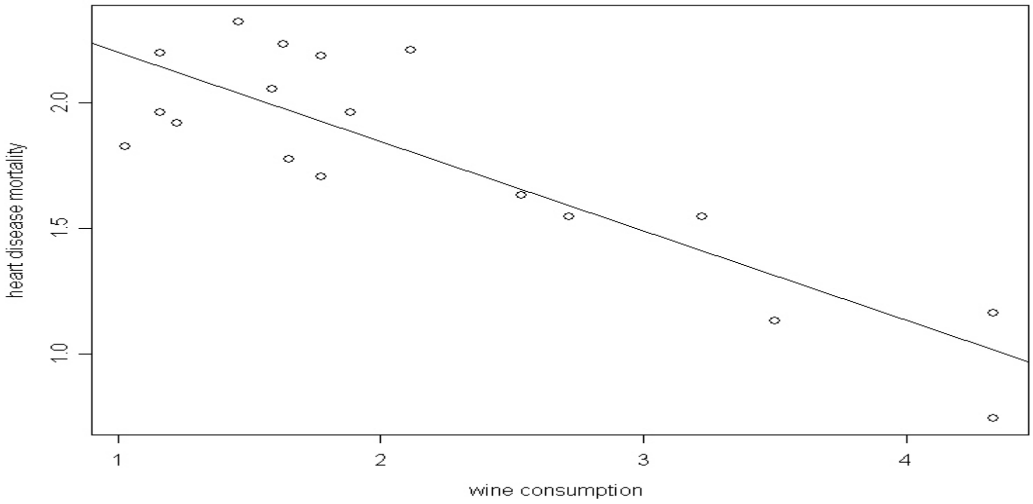
a) Do these data suggest that the heart disease rate is associated with average wine consumption? (You will need to transform the data before fitting the SLR)

The log-log scale provides the best looking straight line fit.

```
winedata<-read.table("ex0823.csv",header=T,sep=",")
wine=winedata$WINE
mortality=winedata$MORTALITY
plot(log(wine),log(mortality),xlab="wine consumption", ylab="heart disease mortality")
wine.reg=lm(log(mortality)~log(wine))
summary(wine.reg)
abline(wine.reg)

Call:
lm(formula = log(mortality) ~ log(wine))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36487 -0.19122  0.01497  0.14485  0.40525 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.55555    0.12690  20.139 8.60e-13 ***
log(wine)   -0.35560    0.05291  -6.721 4.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2285 on 16 degrees of freedom
```

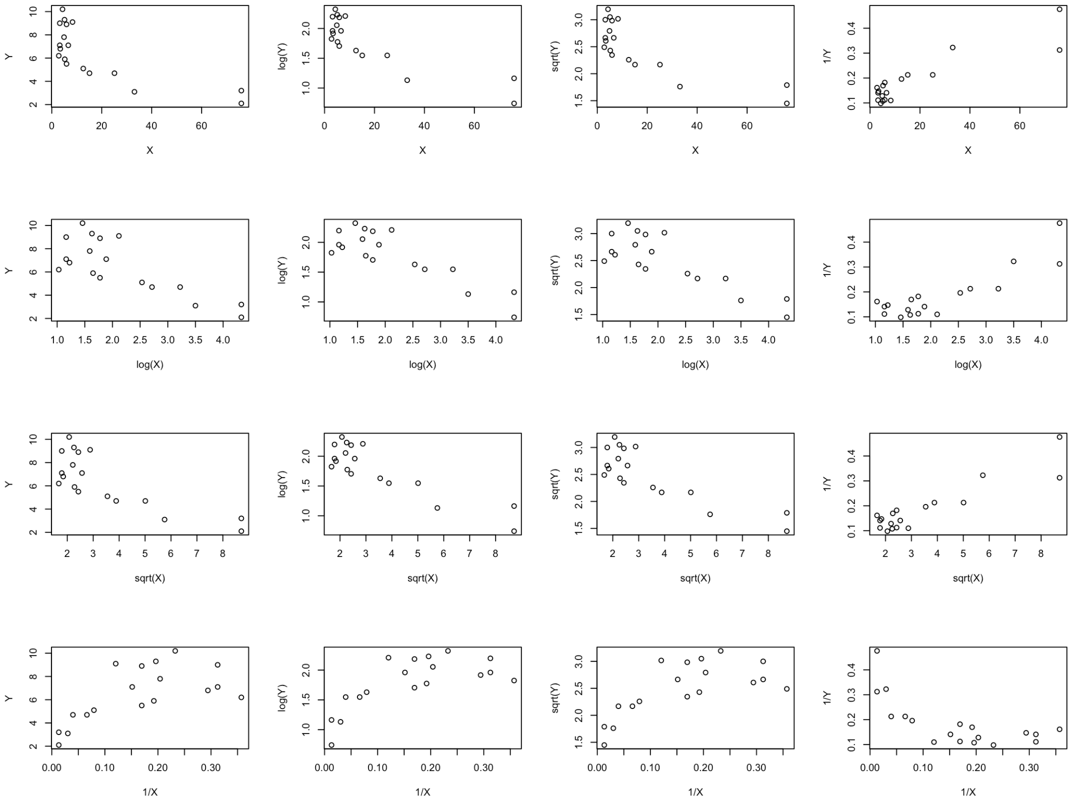
```
Multiple R-squared: 0.7384,     Adjusted R-squared: 0.7221
F-statistic: 45.17 on 1 and 16 DF,  p-value: 4.914e-06
```

There is strong evidence of an association between wine consumption and mortality – the test of null: $\beta_1 = 0$ v's alt: $\beta_1 \neq 0$ has a p-value of <0.01.

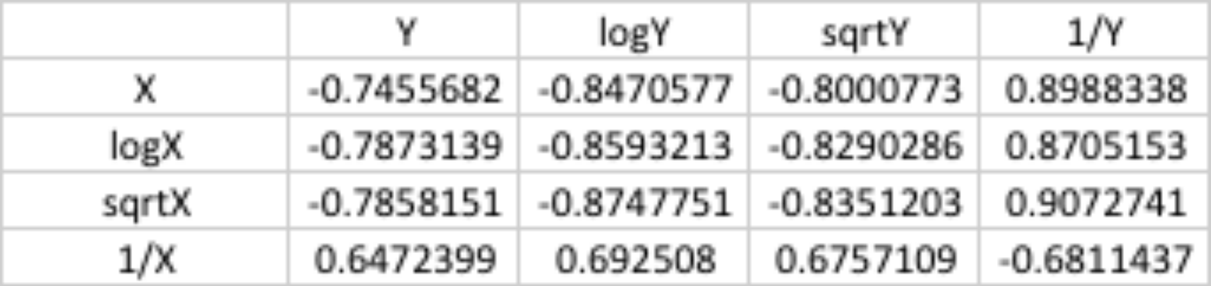


Updates:

In SLR, you can definitely look at the correlation and plots for all possible combinations of transformation to decide which combination of transformation can best fit our data points.



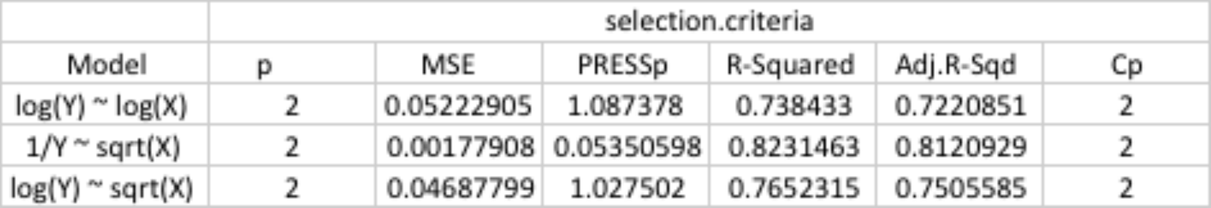
The graph above shows some possible combinations of transformation for this question and below is the associated correlation for each combination of transformation.



	Y	logY	sqrtY	1/Y
X	-0.7455682	-0.8470577	-0.8000773	0.8988338
logX	-0.7873139	-0.8593213	-0.8290286	0.8705153
sqrtX	-0.7858151	-0.8747751	-0.8351203	0.9072741
1/X	0.6472399	0.692508	0.6757109	-0.6811437

Both plots and correlation of possible combinations of transformation suggest that 1/Y~sqrtX and logY~sqrtX are two good alternatives to our logY~logX model.

And we can further confirm this finding by some model selection criteria (which you will see this in model selection lecture soon) as below



	selection.criteria					
Model	p	MSE	PRESSp	R-Squared	Adj.R-Sqd	Cp
log(Y) ~ log(X)	2	0.05222905	1.087378	0.738433	0.7220851	2
1/Y ~ sqrt(X)	2	0.00177908	0.05350598	0.8231463	0.8120929	2
log(Y) ~ sqrt(X)	2	0.04687799	1.027502	0.7652315	0.7505585	2

Notes:

- **p: number of parameters**
- **MSE: mean squared error (measures the difference between the estimator and what is estimated), benchmark: smaller better.**
- **PRESSp: sum of squares for the PRESS residuals, benchmark: smaller better**
- **R-Squared: measured the proportion of response variation "explained" by the regression, benchmark: larger better**
- **Adj.R-Sqd: R-Squared adjusted by degrees of freedom, benchmark: larger better**
- **Cp: estimate sum of the scaled mean squared errors of the fitted values, benchmark: smaller better**

From the statistics above, all three models are appropriate to fit our data points.

Question 2 (revised based on ex 8.17 from Statistical Sleuth)

In a study of the effectiveness of biological control of the exotic weed tansy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragwort. Harvesting the plots the next season, they measured the average dry mass of ragwort remaining (grams/plant) and the flea beetle load (beetles/gram of ragwort mass) to see if the ragwort plants in the plots with high beetle loads were smaller as a result of herbivory by the beetles. The data is contained in the file "ex0817.csv".

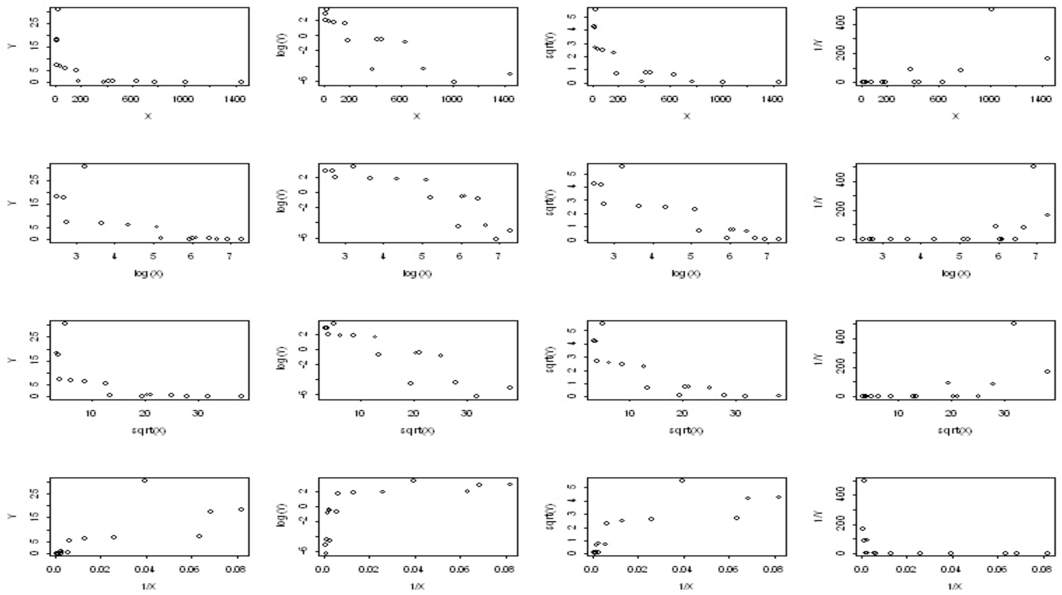
a) Use scatterplots of the raw data, along with trial and error, to find a transformation of Y = Ragwort dry mass and X = Flea beetle load that will be suitable for a SLR. Try the following transformations: log, square root, and reciprocal.

To create the 16 plots covering all transformation combinations use the following commands:

```
Ymat = cbind(Y, log(Y), sqrt(Y), 1/Y)
Xmat = cbind(X, log(X), sqrt(X), 1/X)
Ynames=c(“Y”, “log(Y)”, “sqrt(Y)”, “1/Y”)
Xnames=c(“X”, “log(X)”, “sqrt(X)”, “1/X”)


par(mfrow=c(4,4))

for(i in 1:4) {
    for(j in 1:4) {
                 plot(Xmat[ , i], Ymat[ , j], xlab=Xnames[i], ylab=Ynames[j])
}
}
```

```
weeddata<-read.table("ex0817.csv",header=T,sep=",")
Y<-weeddata$MASS
X<-weeddata$LOAD
Ymat = cbind(Y, log(Y), sqrt(Y), 1/Y)
Xmat = cbind(X, log(X), sqrt(X), 1/X)
Ynames=c("Y", "log(Y)","sqrt(Y)","1/Y")
Xnames=c("X", "log(X)","sqrt(X)","1/X")
par(mfrow=c(4,4))
for(i in 1:4) {
for(j in 1:4) {
plot(Xmat[ , i], Ymat[ , j], xlab=Xnames[i], ylab=Ynames[j])
}
}
```

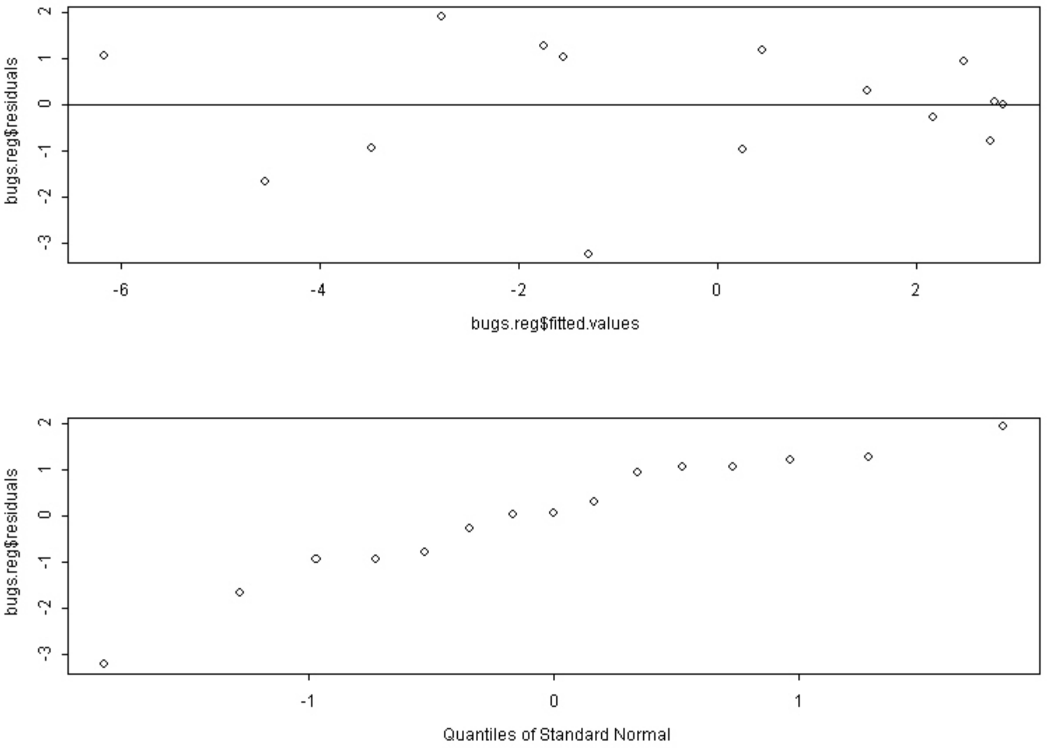
Note: investigate the functions "pairs" for an alternative way of producing pairwise scatterplots.



The plot that is most easily described by a straight line has log(mass) versus sqrt(load). We would fit the regression using log(mass) as the response and sqrt(load) as the explanatory variable).

b) Use a residual plot to check whether the transformations seem appropriate.

```
bugs.reg=lm(log(weeddata$MASS)~sqrt(weeddata$LOAD))
par(mfrow=c(2,1))
plot(bugs.reg$fitted.values,bugs.reg$residuals)
abline(h=0)
qqnorm(bugs.reg$residuals)
```



The residual plot looks satisfactory. For good measure a normal probability plot is also shown. The normal probability plot roughly follows a straight line, suggesting the residuals have a normal distribution. These two plots indicate that the transformations are appropriate.

Question 3 (revised based on ex 7.30 from “The Statistical Sleuth”)
Studies over the last two decades have shown that activity can effect the reorganisation of the human central nervous system. For example, it is known that the part of the brain associated with activity of a finger or limb is taken over for other purposes in individuals whose limb or finger has been lost. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of nine string players (six violinists, two cellists, and one guitarist) and six controls who had never played a musical instrument, when the thumb and fifth finger of the left hand were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers of their left hand extensively, might show different behaviour in the brain – as a result of this extensive physical activity – than individuals who did not play stringed instruments. The file “violin.csv” contains data on the neuron activity index from the MSI and number of years that the individual had been playing a stringed instrument (zero for the controls).

a). *Is the neuron activity different in the stringed musicians and the controls?*

You could answer this question in two ways.

1. **Fit a SLR with an indicator variable for the controls**.

```
violin=read.table("violin.csv",header=T,sep=",")
names(violin)
years=violin$years
act=violin$activity
Iyears=ifelse(years==0,1,0)
diff.reg=lm(act~Iyears)
summary(diff.reg)

    Call:
    lm(formula = act ~ Iyears)

    Residuals:
       Min     1Q Median     3Q    Max
    -9.111 -3.556  1.000  3.694  5.889

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)   20.611      1.534  13.434 5.33e-09 ***
    Iyears       -12.611      2.426  -5.199 0.000171 ***
    ---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Residual standard error: 4.603 on 13 degrees of freedom
    Multiple R-squared: 0.6752,     Adjusted R-squared: 0.6502
    F-statistic: 27.03 on 1 and 13 DF,  p-value: 0.0001714
```

We can then test whether β_1=0. The test statistic is -5.19 with a corresponding p-value of 0.0002. We reject the null and conclude that neuron activity is different in the stringed musicians.

2. **Using a two-sample t-test.**

a) Assuming equal variances

```
v1=var(act[Iyears==1])
v2=var(act[Iyears==0])
m1=mean(act[Iyears==1])
m2=mean(act[Iyears==0])
n1=length(act[Iyears==1])
n2=length(act[Iyears==0])
Spooled=sqrt(((n1-1)*v1+(n2-1)*v2)/(n1+n2-2))
SE=Spooled*sqrt((1/n1)+(1/n2))
Tstat=(m1-m2)/SE
Tstat
[1] -5.1988
```

This gives the same test-statistic and hence p-value as the SLR above.

b) Assuming unequal variances

```
(m1-m2)/(sqrt(v1/6  + v2/9))
[1] -6.067047
```

This gives a slightly different test-statistic. The corresponding p-value is 0.00004.

```
> 2 * pt(-6.1, 13)
[1] 0.00003780384
```

b). *Is the amount of neuron activity associated with the number of years the individual had been playing the instrument?*

```
string.reg=lm(act~years)
summary(string.reg)

Call:
lm(formula = act ~ years)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8644 -2.3730  0.1614  2.3713  4.6471 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.3873     1.1149   7.523 4.35e-06 ***
years         0.9971     0.1110   8.980 6.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.009 on 13 degrees of freedom
Multiple R-squared: 0.8612,     Adjusted R-squared: 0.8505 
F-statistic: 80.63 on 1 and 13 DF,  p-value: 6.178e-07
```

We can then test whether $\beta_1=0$. The test statistic for this test is 9, implying that we reject the null. We estimate that the mean neuronal index increases by 1.0 unit for each one-year increase in musical activity.

Question 4 (revised based on ex 8.25 "The Statistical Sleuth")

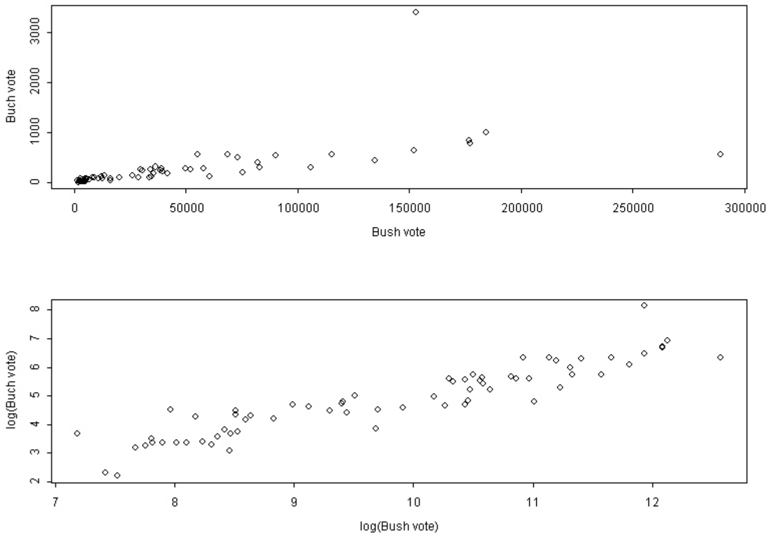
The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns were counted on election night it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was carried by the democrat nominee, Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that the Republican nominee George W. Bush, had carried Florida and won the presidency.

Gore called Bush to concede. While on route to his concession speech, though, the Florida count changed rapidly in his favour. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1738 and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush's lead to be less than 400.

Meanwhile, angry Democrat voters in Palm Beach County complained that a confusing "butterfly" lay-out ballet caused them to accidentally vote for the Reform Party candidate Pat Buchannan instead of Gore. The ballot, shown below, listed presidential candidates on both a left-hand and right-hand page. Voters were to register their vote by punching the circle corresponding to their choice, from the column of circles between the pages. It was suspected that since Bush's name was listed first on the left-hand page, Bush voters likely selected the first circle. Since Gore's name was listed second on the left-hand page, many voters – who already new who they wished to vote for – did not bother examining the right-hand side of the ballet and consequently selected the second circle in the column; the one actually corresponding to Buchannan. Two pieces of evidence supported this claim: Buchannan had an unusually high percentage of the vote in that county, and an unusually large number of ballots (19000) were discarded because voters had marked two circles (possibly inadvertently voting for Buchannan and then trying to correct the mistake by then voting for Gore.)

a). *Produce plots of the number of Buchannan votes versus the number of Bush votes and another plot for the log of these two variables. Does the log-log transformation appear better for performing a simple linear regression?*

```
vote=read.table("vote.csv",header=T,sep=",")
bush=vote$bush
buch=vote$buch
par(mfrow=c(2,1))
plot(bush,buch,xlab="Bush vote",ylab="Buch vote")
plot(log(bush),log(buch),xlab="log(Bush vote)",ylab="log(Buch vote)")
```



The log-log transformation looks better for use in a SLR.

b). Analyse the data without the Palm Beach County results to obtain an equation for predicting Buchannan votes from Bush votes.

```
vote.reg=lm(log(buch[-67])~log(bush[-67]))
summary(vote.reg)
Call:
lm(formula = log(buch[-67]) ~ log(bush[-67]))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.95631 -0.21236  0.02503  0.28102  1.02056 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -2.34149    0.35442  -6.607 9.07e-09 ***
log(bush[-67])  0.73096    0.03597  20.323  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.4198 on 64 degrees of freedom
Multiple R-squared: 0.8658,     Adjusted R-squared: 0.8637 
F-statistic:   413 on 1 and 64 DF,  p-value: < 2.2e-16
```

The fitted regression line is:

$$\hat{\mu}(\text{log(Buch vote)})=\text{-}2.34+0.73\times\text{log(Bush vote)}$$

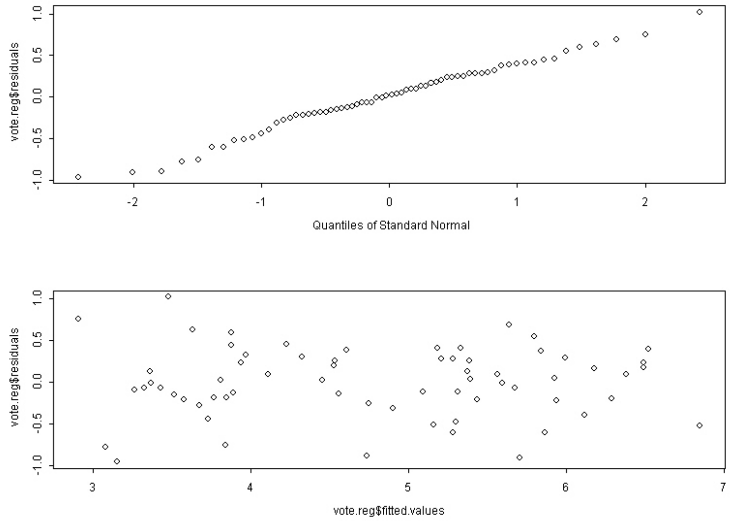
so with Y=Buch vote and X=Bush vote, we have that for a given Bush vote (X_0) that our estimate of the Buch vote ($\hat{Y}$) is:

$$\log(\hat{Y})=\text{-}2.34+0.73\times\log(X_0)$$

$$\Rightarrow \hat{Y}=\exp(\text{-}2.34+0.73\times\log(X_0))$$

c). *Use the residual and normal probability plots from the regression in (b) to check the adequacy of the model.*

```
par(mfrow=c(2,1))
qqnorm(vote.reg$residuals)
plot(vote.reg$fitted.values,vote.reg$residuals)
```



These plots are reasonable, suggesting that SLR model is appropriate.

d). *Obtain a 95% prediction interval for the number of Buchannan votes in Palm Beach – assuming the relationship is the same in this county as the others.*

The number of Bush votes in Palm Beach was 152846. Using the model from b) we predict that there would be $\exp(\text{-}2.34+0.73\times\log(152846))\text{=}586$ votes for Buchannan. A 95% PI interval is given by:

$$\exp\{(\text{-}2.34\text{+}0.73\times 11.94)\pm 2\times 0.42\sqrt{1+\frac{1}{66}+\frac{(11.94-9.78)^2}{(65)1.46}}\}$$

$$=(252,1390)$$

Please try to use the R codes in Lecture Notes 2 to verify this result.

e). *Comment on the result in (d) given that Buchannan actually received 3407 votes in Palm Beach County.*

The observed number of votes for Buchannan is way outside the upper limit of the 95% PI. This is would make us suspect that something was different in Palm Beach.