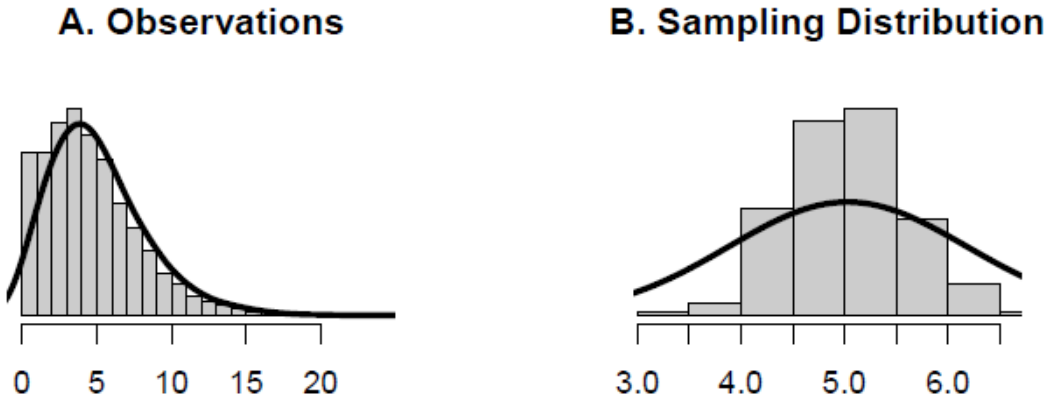
Illapani_Final_Project_IS607

Srini Illapani

December 14, 2015

Part I

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.



a. Describe the two distributions.

The distribution for both A and B appear symmetric, bell shaped and uni-modal although not perfectly normal. Among the both, A is skewed more toward right and B seems to be uniformly distribited compared to A.

b. Explain why the means of these two distributions are similar but the standard deviations are not.

The means are same for both distrubutions because they are are the same population. The SD is different beacuse the distribution range is different for both, figure A has distribution ranging from 0 to 20 and figure B has distribution ranging from 3 to 7 hence the difference in SD.

c. What is the statistical principal that describes this phenomenon?

The central limit theorem. WHich states that given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

Part II

Consider the four datasets, each with two columns (x and y), provided below.
options(digits=2)

```
data1 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))

data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))

data3 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))

data4 <- data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8),
y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately).

```
data1_mean <- c(mean(data1$x), mean(data1$y))
data2_mean <- c(mean(data2$x), mean(data2$y))
data3_mean <- c(mean(data3$x), mean(data3$y))
data4_mean <- c(mean(data4$x), mean(data4$y))

round(data1_mean,2)

## [1] 9.0 7.5

round(data2_mean,2)

## [1] 9.0 7.5

round(data3_mean,2)

## [1] 9.0 7.5

round(data4_mean,2)

## [1] 9.0 7.5
```

b. The median (for x and y separately).

```
data1_median <- c(median(data1$x), median(data1$y))
data2_median <- c(median(data2$x), median(data2$y))
data3_median <- c(median(data3$x), median(data3$y))
data4_median <- c(median(data4$x), median(data4$y))

round(data1_median,2)

## [1] 9.00 7.58
```

```
round(data2_median,2)
```

```
## [1] 9.00 8.14
```

```
round(data3_median,2)
```

```
## [1] 9.00 7.11
```

```
round(data4_median,2)
```

```
## [1] 8.00 7.04
```

c. The standard deviation (for x and y separately).

```
data1_sd <- c(sd(data1$x), sd(data1$y))
data2_sd <- c(sd(data2$x), sd(data2$y))
data3_sd <- c(sd(data3$x), sd(data3$y))
data4_sd <- c(sd(data4$x), sd(data4$y))

round(data1_sd,2)
```

```
## [1] 3.32 2.03
```

```
round(data2_sd,2)
```

```
## [1] 3.32 2.03
```

```
round(data3_sd,2)
```

```
## [1] 3.32 2.03
```

```
round(data4_sd,2)
```

```
## [1] 3.32 2.03
```

For each x and y pair, calculate (also to two decimal places):

d. The correlation.

```
data1_cor <- round(cor(data1$x, data1$y),2)
data2_cor <- round(cor(data2$x, data2$y),2)
data3_cor <- round(cor(data3$x, data3$y),2)
data4_cor <- round(cor(data4$x, data4$y),2)

data1_cor
```

```
## [1] 0.82
```

```
data2_cor
```

```
## [1] 0.82
```

```
data3_cor
```

```
## [1] 0.82
```

```
data4_cor
```

```
## [1] 0.82
```

e. Linear regression equation.

```
data1_lr <- lm(x~y, data = data1)
data2_lr <- lm(x~y, data = data2)
data3_lr <- lm(x~y, data = data3)
data4_lr <- lm(x~y, data = data4)
```

```
data1_lr
```

```
##
## Call:
## lm(formula = x ~ y, data = data1)
##
## Coefficients:
## (Intercept)            y
##     -0.9975       1.3328
```

```
data2_lr
```

```
##
## Call:
## lm(formula = x ~ y, data = data2)
##
## Coefficients:
## (Intercept)            y
##     -0.9948       1.3325
```

```
data3_lr
```

```
##
## Call:
## lm(formula = x ~ y, data = data3)
##
## Coefficients:
## (Intercept)            y
##      -1.000        1.333
```

```
data4_lr
```

```
##
## Call:
## lm(formula = x ~ y, data = data4)
##
## Coefficients:
## (Intercept)            y
##      -1.004        1.334
```

f. R-Squared

```r
data1_r2 <- summary.lm(data1_lr)
data2_r2 <- summary.lm(data2_lr)
data3_r2 <- summary.lm(data3_lr)
data4_r2 <- summary.lm(data4_lr)

data1_r2
```

```
##
## Call:
## lm(formula = x ~ y, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6522 -1.5117 -0.2657  1.2341  3.8946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9975     2.4344  -0.410  0.69156
## y             1.3328     0.3142   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.019 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```r
data2_r2
```

```
##
## Call:
## lm(formula = x ~ y, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8516 -1.4315 -0.3440  0.8467  4.2017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9948     2.4354  -0.408  0.69246
## y             1.3325     0.3144   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.02 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```r
data3_r2
```

```
##
## Call:
```

```
## lm(formula = x ~ y, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9869 -1.3733 -0.0266  1.3200  3.2133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0003     2.4362  -0.411  0.69097
## y             1.3334     0.3145   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.019 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

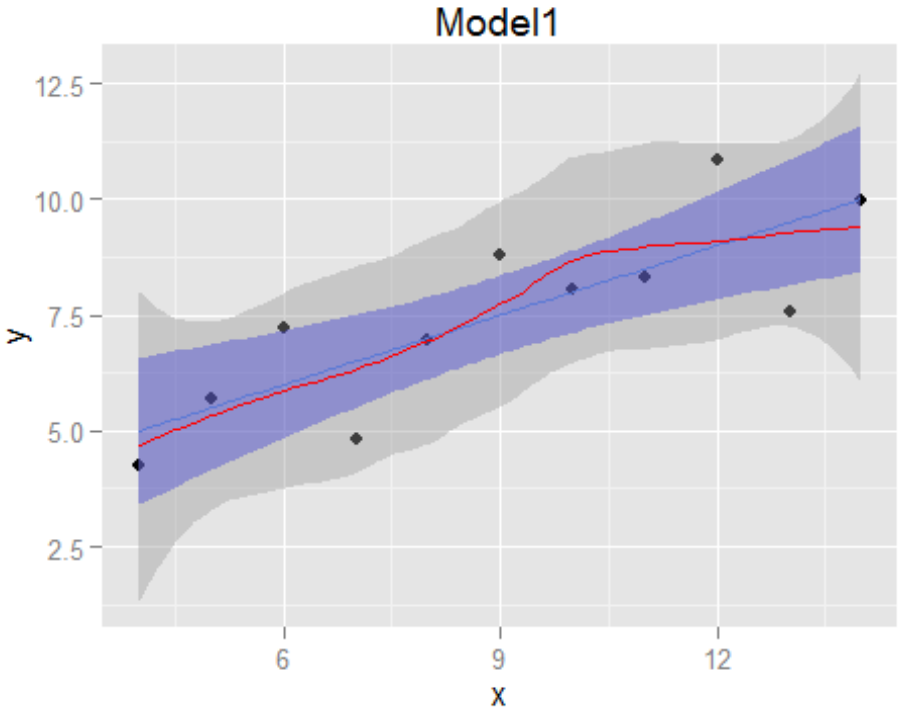
```
data4_r2
```

```
##
## Call:
## lm(formula = x ~ y, data = data4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7859 -1.4122 -0.1853  1.4551  3.3329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0036     2.4349  -0.412  0.68985
## y             1.3337     0.3143   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.018 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

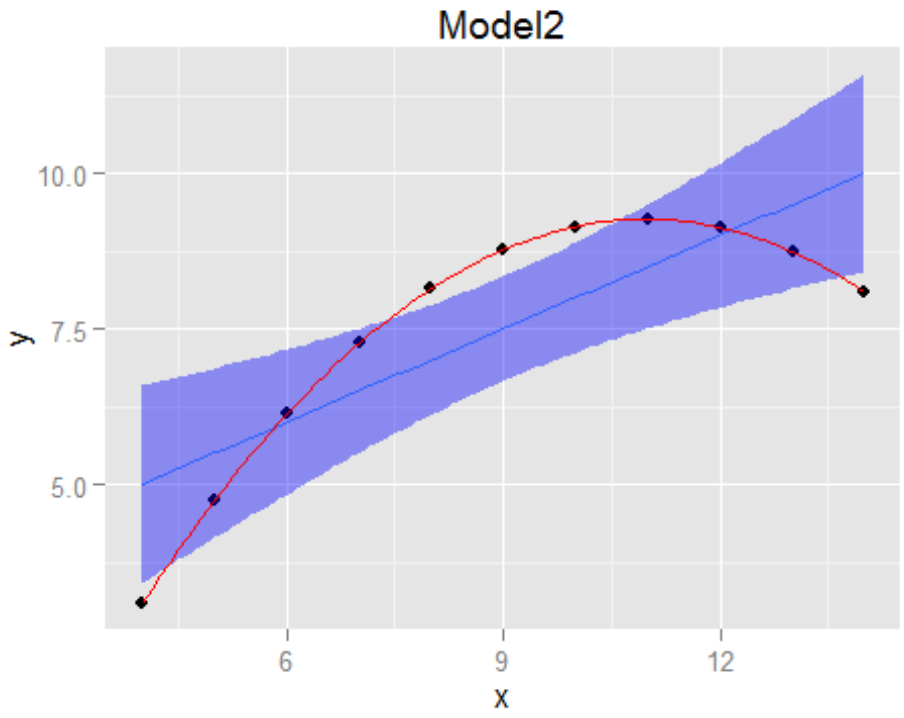
g. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair!

```
library(ggplot2)

plot_data1 <- ggplot(aes(x,y), data = data1) + geom_point() +
              geom_smooth(method='lm',formula=y~x, fill = "blue") +
              stat_smooth(colour="red") + labs(title = "Model1")

plot_data2 <- ggplot(aes(x,y), data = data2) + geom_point() +
              geom_smooth(method='lm',formula=y~x, fill = "blue") +
              stat_smooth(colour="red") + labs(title = "Model2")
```

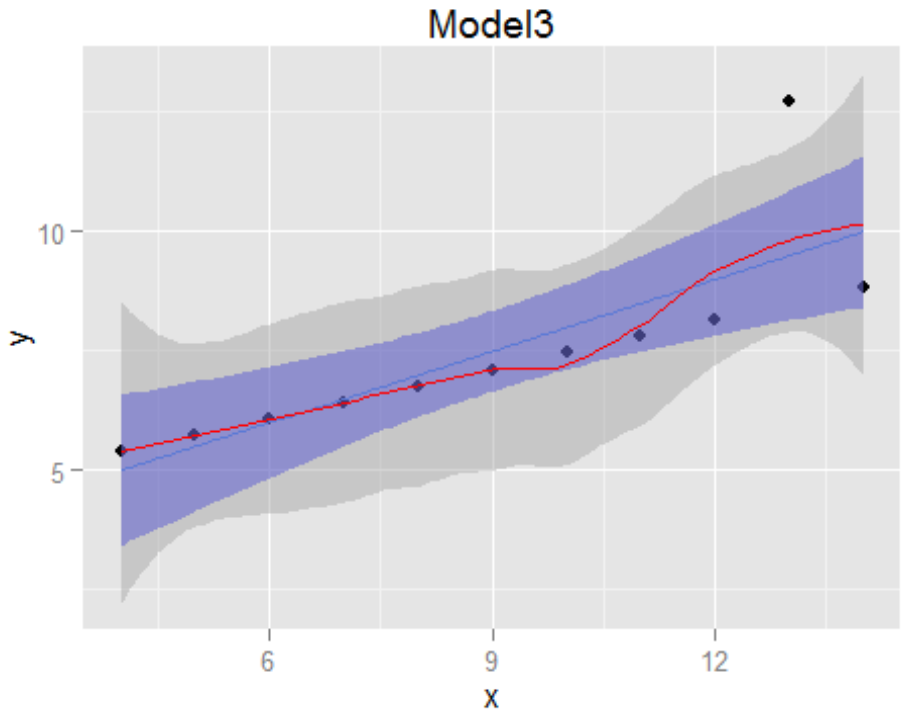
```r
plot_data3 <- ggplot(aes(x,y), data = data3) + geom_point() +
              geom_smooth(method='lm',formula=y~x, fill = "blue") +
              stat_smooth(colour="red") + labs(title = "Model3")

plot_data4 <- ggplot(aes(x,y), data = data4) + geom_point() +
              geom_smooth(method='lm',formula=y~x, fill = "blue") +
              stat_smooth(colour="red") + labs(title = "Model4")

plot_data1
```



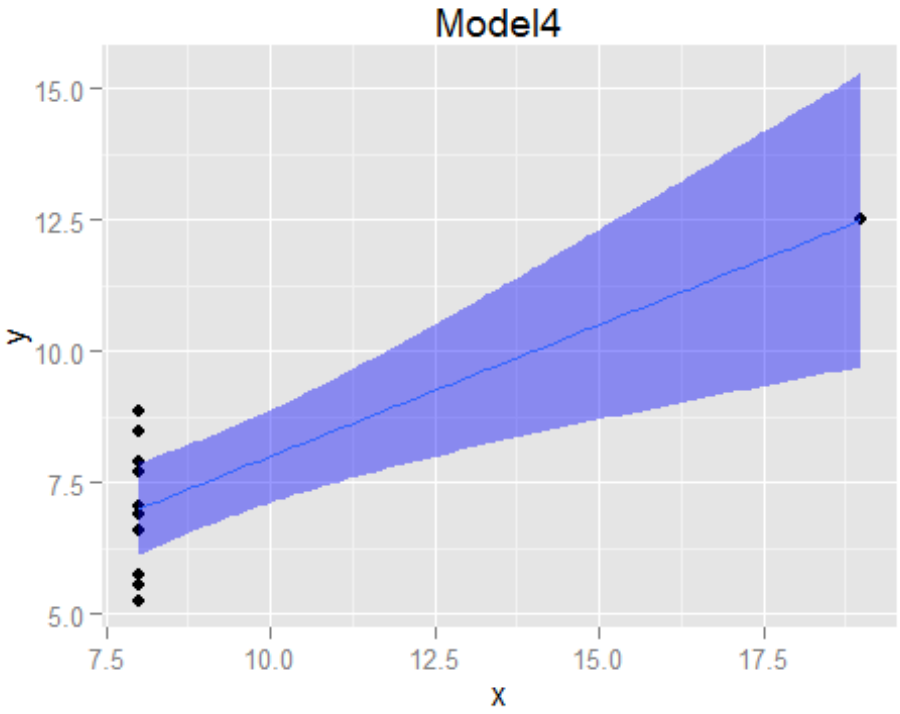

```r
plot_data2
```




```
plot_data3
```




```
plot_data4
```



As one can see from the above plots, not all the data sets exhibit the linear regression model. Model 1 and Model 3 do fit the linear regression as we can see the correlation of x and y variables. Model 2 and Model 4 do not fit the linear regression model.

So it is not appropriate to apply or fit a linear regression for all the models.

h. Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create.

The distribution and quantitative measurements for the four datasets do not give out the major differences we saw between the data sets but once we apply visualization by drawing plots, we see the differences between the data sets. So it is important to bring visualization for the data sets when analyzing data.