DATA 606 Spring 2017 - Final Exam

Ilya Kats

May 19, 2017

Part I

Question 1

Answer: (b) – Variable `daysDrive` is quantitative and discrete as defined since the value is counted in whole days. Variables `car` and `color` are qualitative (even though values of `car` are coded with whole numbers). Variable `gasMonth` is continuous.

Question 2

Answer: (a) – Since the histogram is heavily left skewed, median must be greater than mean. Median of 3.8 is too high (on the histogram median means that bars to the left of it and bars to the right of it when stacked will appear of equal height). This leaves us with (a).

Question 3

Answer: (a) – Only experiments can show causation between events.

Question 4

Answer: (c) – The chi-square test with a large value will have a small p-value, which will cause us to reject the null hypothesis. The null hypothesis in this example would've been that distribution of eye color for each hair color group matches the random distribution for eye color in general. There is an association between natural hair color and eye color.**

Question 5

Answer: (b)

$IQR = Q3 - Q1 = 49.8 - 37 = 12.8$

Outliers will generally lie 1.5 IQR away from Q1 and Q3, so below $Q1 - 1.5 \times IQR = 17.8$ and above $Q3 + 1.5 \times IQR = 69$.

Question 6

Answer: (d) – The median and interquartile range are resistant to outliers, whereas the mean and standard deviation are not.

Question 7

a. Distribution A is unimodal and left-skewed. Distribution B is unimodal, symmetrical and nearly normal.

b. Distribution B represents samples from population A. Each sample is random, so its mean should be similar to population mean. And while the mean of one sample may differ due to natural variations, the mean of large number of samples is very close to the actual population mean. However, distributions A and B represent different things - one is distribution of observations while the other is distribution of sample means across multiple samples. So standard deviation for A illustrates variability of observations and standard deviation for B illustrates variability of sample mean. These two standard deviations are not related, so can differ widely.

c. This phenomenon is described by the *Central Limit Theorem*.

Part II

```
# Initial data
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))

# Create data frame to hold results
stat <- data.frame(Object=c("data1", "data2", "data3", "data4"),
                   Mean_X=c(0, 0, 0, 0), Mean_Y=c(0, 0, 0, 0),
                   Median_X=c(0, 0, 0, 0), Median_Y=c(0, 0, 0, 0),
                   SD_X=c(0, 0, 0, 0), SD_Y=c(0, 0, 0, 0),
                   Correlation=c(0, 0, 0, 0))

# Calculate means
stat$Mean_X[1] <- mean(data1$x)
stat$Mean_X[2] <- mean(data2$x)
stat$Mean_X[3] <- mean(data3$x)
stat$Mean_X[4] <- mean(data4$x)
stat$Mean_Y[1] <- mean(data1$y)
stat$Mean_Y[2] <- mean(data2$y)
stat$Mean_Y[3] <- mean(data3$y)
stat$Mean_Y[4] <- mean(data4$y)

# Calculate medians
stat$Median_X[1] <- median(data1$x)
stat$Median_X[2] <- median(data2$x)
stat$Median_X[3] <- median(data3$x)
stat$Median_X[4] <- median(data4$x)
stat$Median_Y[1] <- median(data1$y)
stat$Median_Y[2] <- median(data2$y)
stat$Median_Y[3] <- median(data3$y)
stat$Median_Y[4] <- median(data4$y)
```

```r
# Calculate standard deviations
stat$SD_X[1] <- sd(data1$x)
stat$SD_X[2] <- sd(data2$x)
stat$SD_X[3] <- sd(data3$x)
stat$SD_X[4] <- sd(data4$x)
stat$SD_Y[1] <- sd(data1$y)
stat$SD_Y[2] <- sd(data2$y)
stat$SD_Y[3] <- sd(data3$y)
stat$SD_Y[4] <- sd(data4$y)
```

The table below lists means, medians and standard deviations for **x** and **y** of all 4 data frames as requested in **parts a, b, and c**.

Object	Mean_X	Mean_Y	Median_X	Median_Y	SD_X	SD_Y
data1	9	7.5	9	7.6	3.3	2
data2	9	7.5	9	8.1	3.3	2
data3	9	7.5	9	7.1	3.3	2
data4	9	7.5	8	7.0	3.3	2

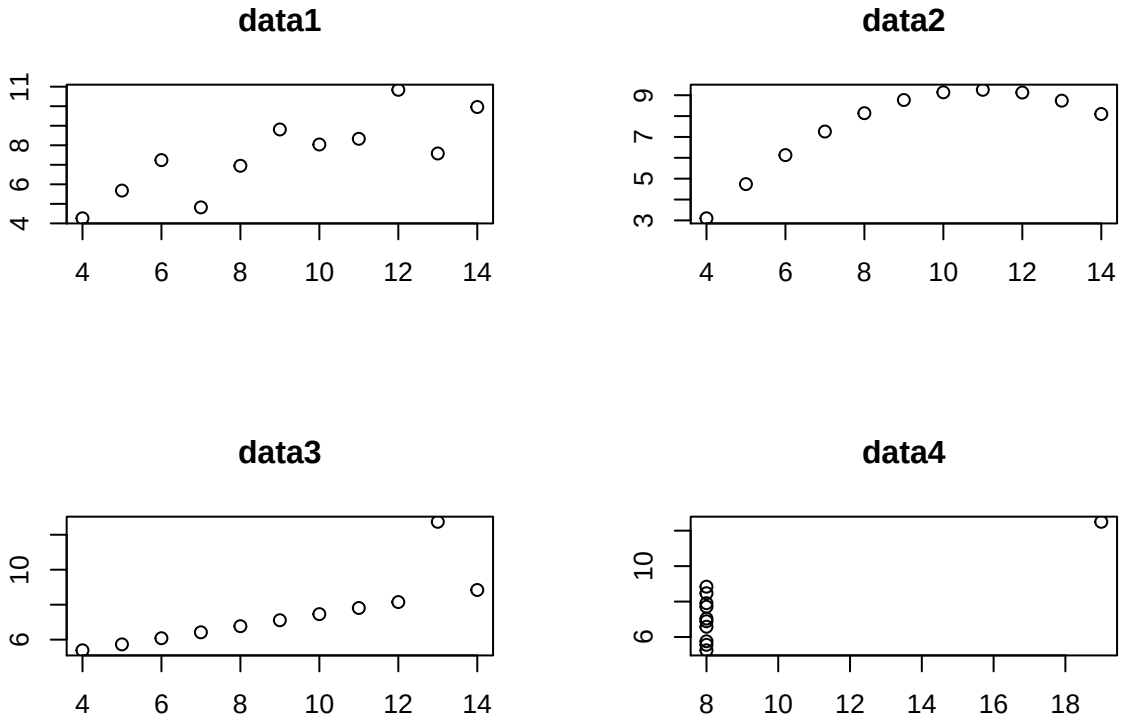
```r
# Calculate correlation
stat$Correlation[1] <- sum((data1$x - stat$Mean_X[1]) * (data1$y - stat$Mean_Y[1])) /
  sqrt(sum((data1$x - stat$Mean_X[1])^2) * sum((data1$y - stat$Mean_Y[1])^2))
stat$Correlation[2] <- sum((data2$x - stat$Mean_X[2]) * (data2$y - stat$Mean_Y[2])) /
  sqrt(sum((data2$x - stat$Mean_X[2])^2) * sum((data2$y - stat$Mean_Y[2])^2))
stat$Correlation[3] <- sum((data3$x - stat$Mean_X[3]) * (data3$y - stat$Mean_Y[3])) /
  sqrt(sum((data3$x - stat$Mean_X[3])^2) * sum((data3$y - stat$Mean_Y[3])^2))
stat$Correlation[4] <- sum((data4$x - stat$Mean_X[4]) * (data4$y - stat$Mean_Y[4])) /
  sqrt(sum((data4$x - stat$Mean_X[4])^2) * sum((data4$y - stat$Mean_Y[4])^2))

# Alternatively, this could've been computated as
# cor(data1$x, data1$y)
```

Correlations for **part d** are as follows:

Object	Correlation
data1	0.82
data2	0.82
data3	0.82
data4	0.82

Basic statistics are very similar for these data sets. And correlations are identical (and at 0.82 it is a fairly strong positive correlation). Plots below illustrate each data set.

```r
par(mfrow=c(2,2))
plot(data1$x, data1$y, xlab = "", ylab = "", main = "data1")
plot(data2$x, data2$y, xlab = "", ylab = "", main = "data2")
plot(data3$x, data3$y, xlab = "", ylab = "", main = "data3")
plot(data4$x, data4$y, xlab = "", ylab = "", main = "data4")
```



I will use R functions to determine linear regression equation.

```
# Build linear regression model for each data set
l1 <- lm(y ~ x, data = data1)
l2 <- lm(y ~ x, data = data2)
l3 <- lm(y ~ x, data = data3)
l4 <- lm(y ~ x, data = data4)

coefficients <- rbind(l1$coefficients, l2$coefficients,
                      l3$coefficients, l4$coefficients)
rownames(coefficients) <- c("data1", "data2", "data3", "data4")

coefficients
```

```
##       (Intercept)   x
## data1           3 0.5
## data2           3 0.5
## data3           3 0.5
## data4           3 0.5
```

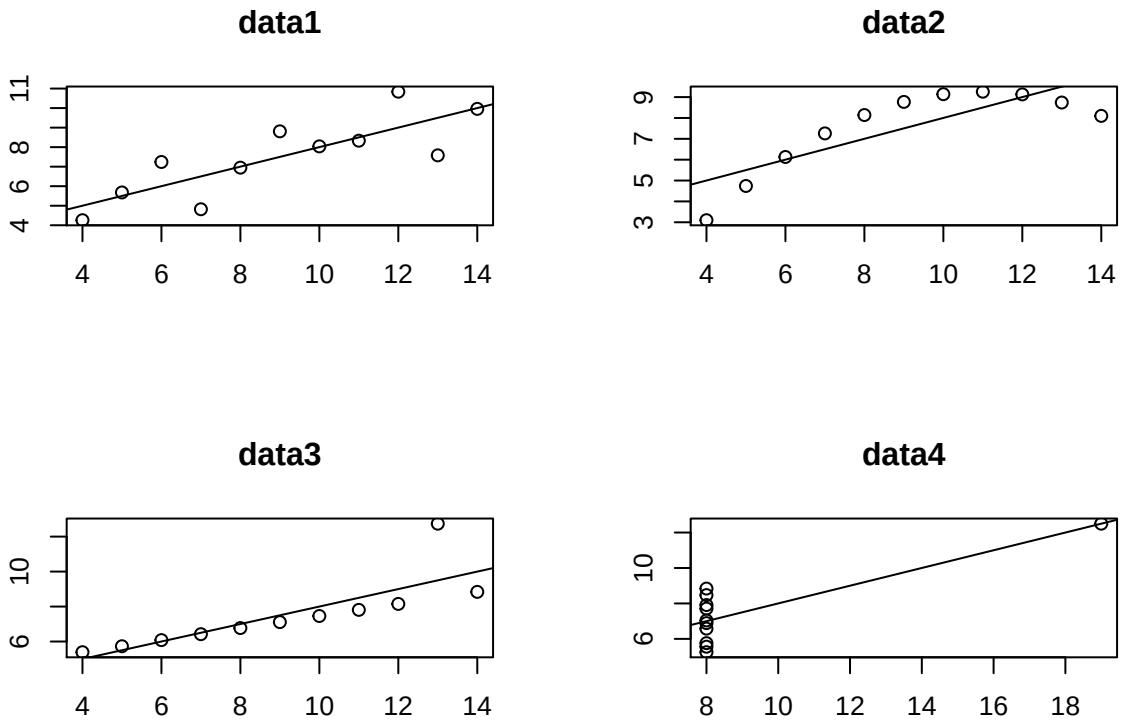
Linear regression equations is the same for all data sets (**part e**):

$y = 3 + 0.5x$

Correlation is $R = 0.82$ for all data sets. So R-squared is $R^2 = 0.67$ for all data sets (**part f**).

The plots below show all data sets with regression lines.

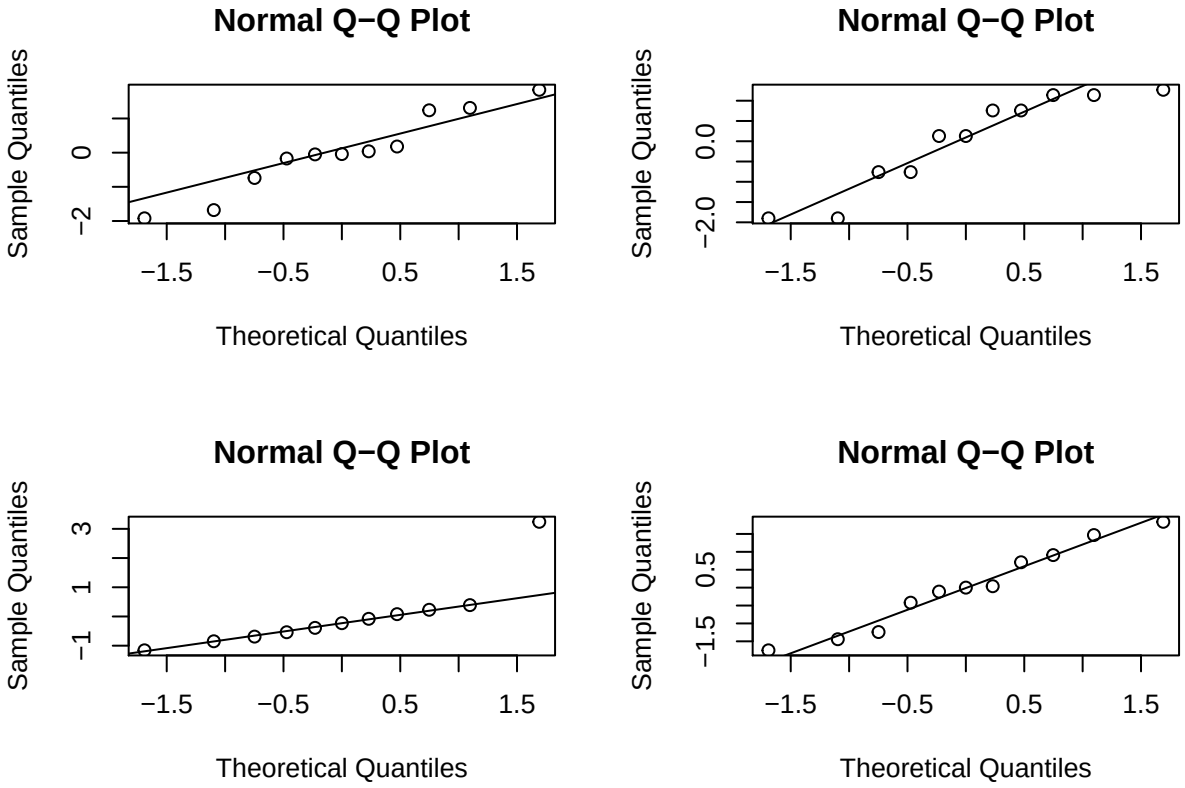
```
par(mfrow=c(2,2))
plot(data1$x, data1$y, xlab = "", ylab = "", main = "data1")
abline(l1)
plot(data2$x, data2$y, xlab = "", ylab = "", main = "data2")
abline(l2)
plot(data3$x, data3$y, xlab = "", ylab = "", main = "data3")
abline(l3)
plot(data4$x, data4$y, xlab = "", ylab = "", main = "data4")
abline(l4)
```



Linearity: Plots above show that only `data1` shows linear trend. `data2` shows relationship between x and y, but it is not linear. `data3` has for the most part a very linear relationship, but has am outlier that may skew the model. Similarly, 'data4' has mostly a linear relationship except for one very extreme outlier that completely skews the model.

Consider normal probability plots of residuals:

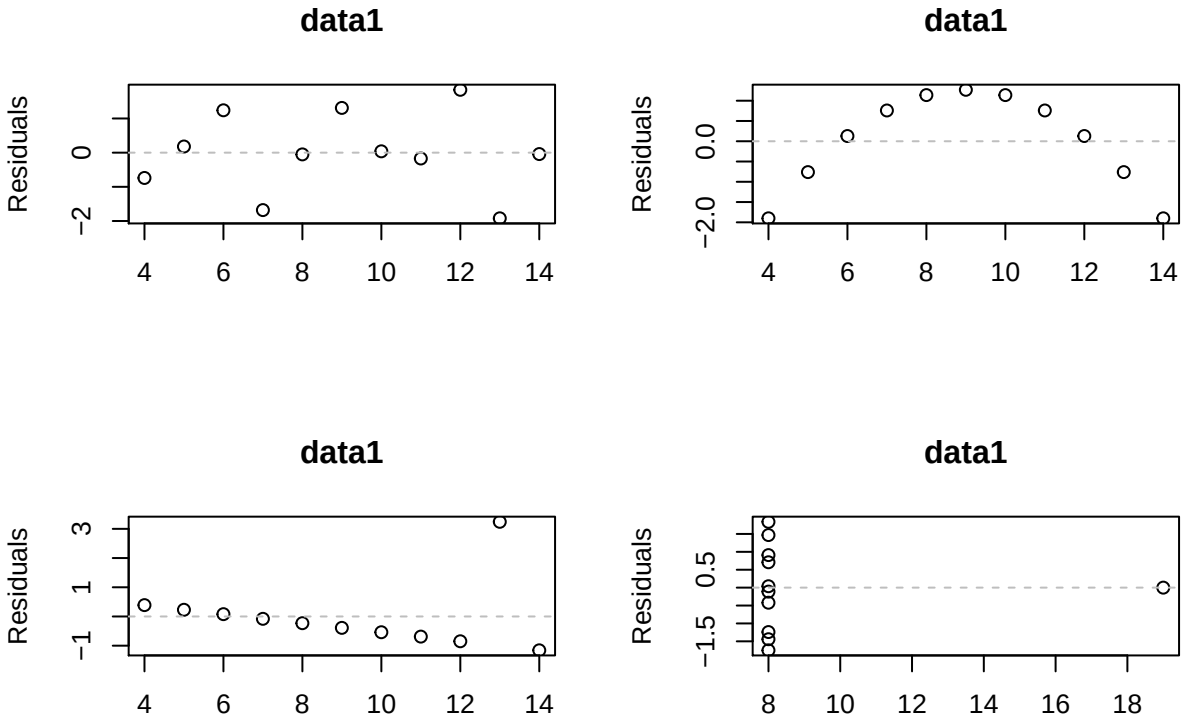
```
par(mfrow=c(2,2))
qqnorm(l1$residuals)
qqline(l1$residuals)
qqnorm(l2$residuals)
qqline(l2$residuals)
qqnorm(l3$residuals)
qqline(l3$residuals)
qqnorm(l4$residuals)
qqline(l4$residuals)
```



Nearly normal residuals: For the most part the distribution of residuals is nearly normal for all data sets with a possible exception of `data3` which has an outlier.

Consider residual plots:

```
par(mfrow=c(2,2))
plot(data1$x, l1$residuals, xlab = "", ylab = "Residuals", main = "data1")
abline(h=0,lty=2,col="grey")
plot(data2$x, l2$residuals, xlab = "", ylab = "Residuals", main = "data1")
abline(h=0,lty=2,col="grey")
plot(data3$x, l3$residuals, xlab = "", ylab = "Residuals", main = "data1")
abline(h=0,lty=2,col="grey")
plot(data4$x, l4$residuals, xlab = "", ylab = "Residuals", main = "data1")
abline(h=0,lty=2,col="grey")
```



Constant variability: Residual plots clearly show that `data2`, `data3`, and `data4` do not have constant variability of residuals. Only `data1` can be considered as satisifying this condition.

Independent observations: We do not know what these data sets represent, so we cannot evaluate the independence of observations.

Based on the analysis above, the linear regression model is only appropriate for `data1` (provided we confirm independence). All other data sets fail linearity and variability conditions.

These data sets have very similar medians and identical means, standard deviations, correlations, R-squared and linear regression equations. Without plotting and checking conditions we might have concluded that the data sets are similar and can be described with a linear regression model with a strong correlation. However, even a simple scatterplot shows right away that the linear regression model is completely inapproriate for some data sets and may need serious checking of conditions for some others. This is why it is critical to visuals the data and check conditions when evaluating models.