Arif Ali
ANLY-511 Prob. Modeling & Stat. Computing
Dec 09, 2015

Homework #11

Exercise 81

Part A

$\mu_0 = 98.6$ and $\mu_a > 98.6$

Part B

```
Sodor.kidstemp = c(98.0, 98.9, 99.0, 98.9, 98.8,
                   98.6, 99.1, 98.9, 98.5, 98.9,
                   98.9, 98.4, 99.0, 99.2, 98.6,
                   98.8, 98.9, 98.7)
```

```
pt((mean(Sodor.kidstemp)-98.6)/
  (sd(Sodor.kidstemp)/sqrt(length(Sodor.kidstemp))),
  df = length(Sodor.kidstemp)-1, lower.tail = F)
## [1] 0.006906586
```

Since the p-value is less than 0.05, we can conlcude that the average child's temperture in Sodor is higher than normal.

extra credit 1

```
Sodor.kidstemp.ci = replicate(10000, mean(sample(Sodor.kidstemp, length(Sodor.kidstemp), replace = T)))
quantile(Sodor.kidstemp.ci, c(0.05))
##       5%
## 98.66667
```

Based on the confidence interval, we can be 95% is between 98.65 and 98.90, which follows the idea from Exercise 81 Part B, where the p-value is highly significiant. We can, again, conclude that the average child's temperture in Sodor is higher than normal.

Exercise 82

Part A

```
prop.test(c(28,13),c(250,250))$conf.int
## [1] 0.008191018 0.111808982
## attr(,"conf.level")
## [1] 0.95
```

We are 95%confident that the difference in infection between the two oxygen groups is between 0.008191018 and 0.111808982.

Part B

Because there is no control group, we cannot determine whether the adding oxygen to the samples result in some sort difference (whether significant or not).

Exercise 83

Part A

Type I error: If after testing the 15 households, we get a P-value given an α such that we reject the null hypothesis in favor of the alternative hypothesis $\mu > 10$. However, the aresenic level is still $\mu = 10$. Thus, a significant amount of money could be invested to clean up arsen, even though it doesn't pose a series problem. Type II error, the sample tests show that the level aren't elevated. Even though there is an arsenic level problem, it's dismissed. People may die due to this situation.

Part B

The T-distribution will look approximately the same as using the standard normal distribution; but it will not neccesarily be the same, hence the approximation. The tails will move further away from zero than with a standard normal distribution.

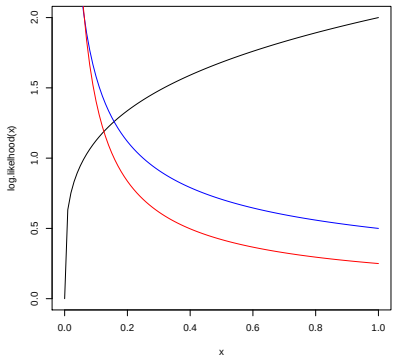
extra credit 2

Part A

$H_0 : \theta = 1/2$ and $H_A : \theta = 1/4$ and $f(x;\theta) = \theta x^{\theta-1}$

$$L(\theta) = \frac{1/2x^{-1/2}}{1/4x^{-3/4}} = 2x^{-1/2+3/4} = 2x^{1/4} \tag{1}$$

```
log.likelihood = function(x){2*x^(1/4)}
curve(log.likelihood)

h0 = function(x){1/2*x^(-1/2)}
curve(h0, add = T, col = "blue")

ha = function(x){1/4*x^(-3/4)}
curve(ha, add = T, col = "red")
```



$$\int_0^c \frac{1}{2}x^{-\frac{1}{2}}dx = \Big|_0^c \sqrt{x} = 0.05 \implies c = (0.05)^2 = 0.0025 \tag{2}$$

Part B

$$\int_0^{0.0025} \frac{1}{4} x^{-\frac{3}{4}} dx = \Big|_0^{0.0025} \sqrt[4]{x} = 0.223607 \tag{3}$$

Exercise 84

Part A

$\mu_y = 94$ and $\sigma_y = 15$, $\mu_x = 46$ and $\sigma_x = 7$, $\rho = 0.75$

$$weight = 1.607143 * height + 20.07142 \tag{4}$$

Part B

```
1.607143*(12*5) + 20.07142
## [1] 116.5
```

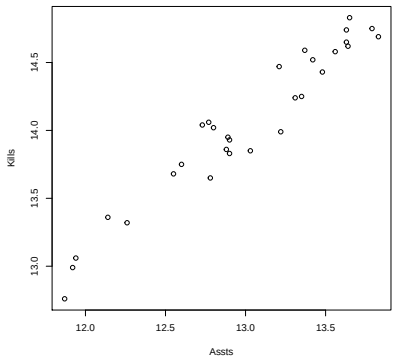
Part C

```
0.75*0.75
## [1] 0.5625
```

extra credit 3

Part A

```
Volleyball2009 <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/Volley
plot(Kills~Assts, data = Volleyball2009)
```



There seems to be a positive relationship between Kills and Assts. This is based on a pattern that could be fitted to a line with a positive slope.

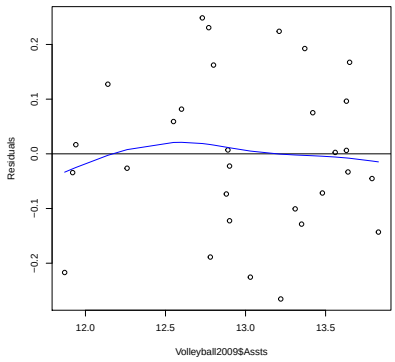
Part B

```
part_b = lm(Kills~Assts, data = Volleyball2009)
part_b$coefficients
## (Intercept)       Assts
##   1.7362551   0.9469872
summary(part_b)$r.squared
## [1] 0.9367418
```

The slope is positive and close to one, confirming my assumptions from the scatter plot. The R Square is about very high, so a significiant amount of the variability in kills is explained via the regression.

Part C

```
plot(Volleyball2009$Assts, resid(part_b), ylab = "Residuals")
 abline(h=0)

lines(smooth.spline(Volleyball2009$Assts, resid(part_b), df = 3),
col="blue")
```

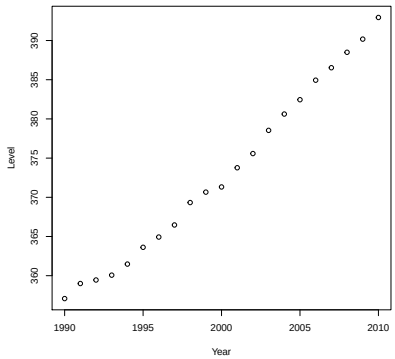


Based on the residuals, it seems like the straight line method is appropiate given the problem.

Exercise 85

Part A

```
Maunaloa <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/Maunaloa.cs

plot(Level~Year, data = Maunaloa)
```



There seems to be a postive relationship between the Level of CO_2 and Year. This due to a positive slope of a line that could be fitted to the scatter plot. The intercept will be positive as well based on the range of Levels.

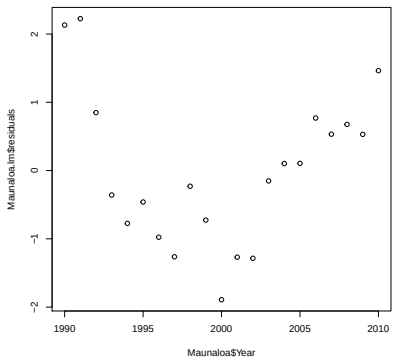
Part B

```
Maunaloa.lm = lm(Level~Year, data = Maunaloa)
Maunaloa.lm$coefficients
##  (Intercept)         Year
## -3279.592814     1.826403
```

$$Level = 1.826403 * Year - 3279.592814 \tag{5}$$

Part C

```
plot(Maunaloa$Year,Maunaloa.lm$residuals)
```



The residuals are very scattered, so the straight line model may not be the most appropiate measure. (The back of the book indicates it's probably serial correlation), which states that past measures affect future measurement, which makes sense because how CO_2 levels are measured.

Exercise 86

Part A

Since X_i $Bern(p)$ then Y $Bin(12, p)$ So the Probability of the type I error is: $P(RejectH_0|H_0true) = P(\hat{p} < 0.3|p = 0.3 = P(k = 0|p = 0.3) + P(k = 1|p = 0.3) = (0.3)^0 * (1 - 0.3)^{12} + \begin{pmatrix} 12 \\ 1 \end{pmatrix} * (0.3)^1 * (1 - 0.3)^{11} = 0.0850250$

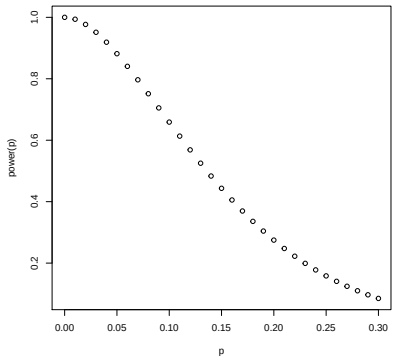
Part B

$1-\beta = P(RejectH_0|H_Atrue) = P(H_0 \neq p|\hat{p} < p) = 1-P(H_0 = p|\hat{p} < p) = ((p)^0*(1-p)^{12}+12*(p)^1*(1-p)^{11})$

```
power = function(p){
  ((p)^0*(1-p)^12+12*(p)^1*(1-p)^11)
}
```

Part C

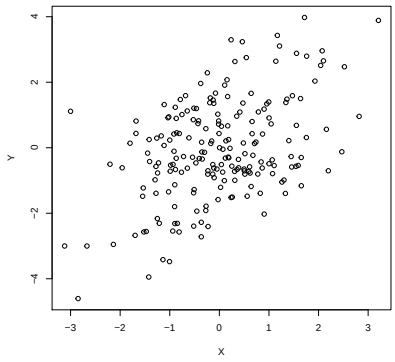
```
p = seq(0,.3,0.01)
plot(p, power(p))
```



Exercise 87

Part A

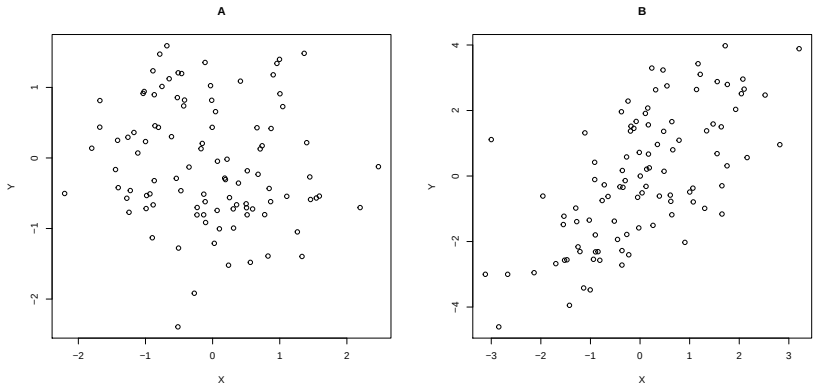
```
corrExerciseA <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/corrEx
plot(Y~X, data = corrExerciseA)
```




```
cor(corrExerciseA$X, corrExerciseA$Y)
## [1] 0.4550343
```

Part B

```
plot(Y~X,
     data = corrExerciseA[corrExerciseA$Z=="A",],
     main = "A")

plot(Y~X,
     data = corrExerciseA[corrExerciseA$Z=="B",],
     main = "B")
```



Plot A is trending downward, so there must be a negative relation between X and Y in subset A. Plot B indicates a more positive relationsip between X and Y in subset B compared to subset A or overall.

Part C

```
print("A")
## [1] "A"
cor(corrExerciseA[corrExerciseA$Z=="A","X"],
    corrExerciseA[corrExerciseA$Z=="A","Y"])
## [1] -0.1335408
print("B")
```

```
## [1] "B"
cor(corrExerciseA[corrExerciseA$Z=="B","X"],
    corrExerciseA[corrExerciseA$Z=="B","Y"])
## [1] 0.65303
```

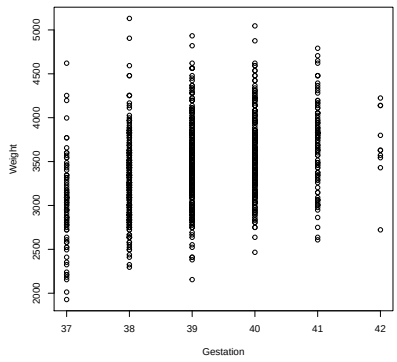
Part D

Based on the correlations, it seems that a very large positive correlation within a dataset will help mask the fact that other subsets have negative correlations.

Exercise 88

Part A (17)

```
NCBirths2004 <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/NCBirths
plot(Weight~Gestation, data = NCBirths2004)
```




```
cor(NCBirths2004$Weight, NCBirths2004$Gestation)
## [1] 0.3486057
```

Part B (17)

```
part_b = lm(Weight~Gestation, data = NCBirths2004)
part_b$coefficients
## (Intercept)   Gestation
##  -2379.6896    148.9954
```

$$Weight = 148.9954 * Gestation - 2379.6896 \tag{6}$$

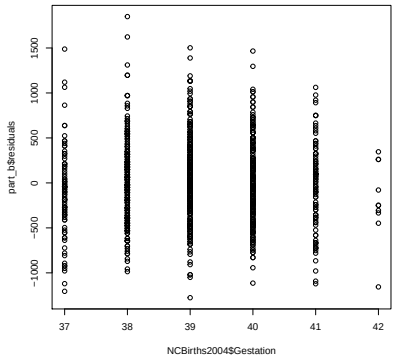
Part C (17)

```
summary(part_b)$r.squared
## [1] 0.1215259
```

The slope helps indicate a postive relationship between Gestation and Weight. However, based on the R^2, the fact that it's small means that most of the data isn't close fitted to the line, which makes sense from the plot where the range of Weights by Gestation is spread out.

Part D (17)

```
plot(NCBirths2004$Gestation, part_b$residuals)
```



The residual plots looks significiantly like the plot from Part A. We need to look as each Gestation individually, because there is wide weight ranges for each period.

Part A (18)

```
summary(part_b)
##
## Call:
## lm(formula = Weight ~ Gestation, data = NCBirths2004)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1276.13  -312.13   -22.13   267.88  1848.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2379.69     493.99  -4.817 1.68e-06 ***
## Gestation     149.00      12.62  11.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 457.4 on 1007 degrees of freedom
## Multiple R-squared:  0.1215,Adjusted R-squared:  0.1207
## F-statistic: 139.3 on 1 and 1007 DF,  p-value: < 2.2e-16
```

Estimate of σ is 457.4.

Part B (18)

```
summary(part_b)$coefficients[,1][2] +
  1.96*c(-1,1)*summary(part_b)$coefficients[,2][2]
## [1] 124.2529 173.7379
```

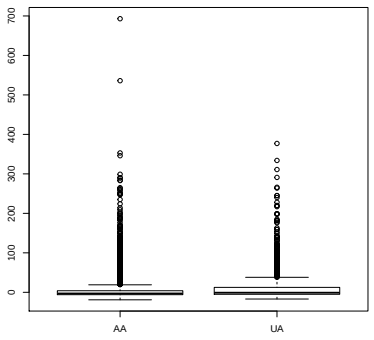
extra credit 4

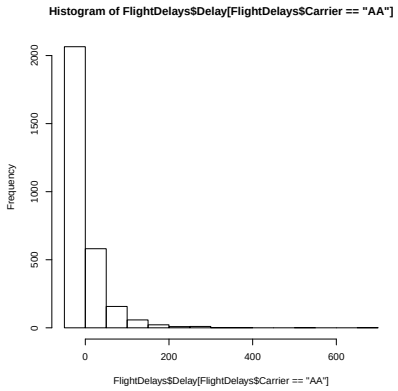
```
Alelager <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/Alelager.csv
cor(Alelager$Alcohol, Alelager$Calories)
## [1] 0.5371458
Alesampler = function(){
  index = sample(1:nrow(Alelager), nrow(Alelager), replace = T)
  Alelager.bool = Alelager[index, ]
  cor(Alelager.bool$Alcohol, Alelager.bool$Calories)
}


Alelager.bootstrap.ci = replicate(10000, Alesampler())

quantile(Alelager.bootstrap.ci, c(0.025, .975))
##      2.5%     97.5%
## 0.3699625 0.7498074
```

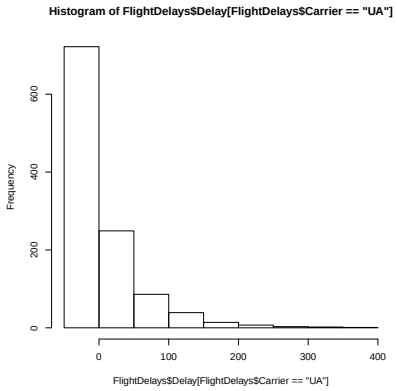
extra credit 5

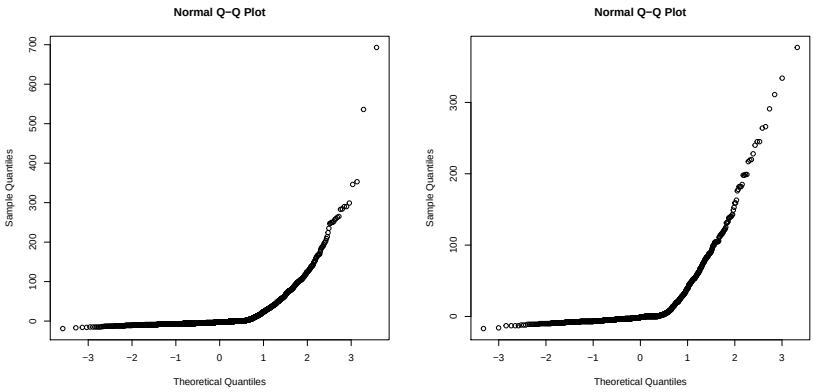
Part A

```
FlightDelays <- read.csv("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/ChiharaHesterberg/FlightDe
FlightDelays$Carrier = as.character(FlightDelays$Carrier)
boxplot(Delay~Carrier, data = FlightDelays)
hist(FlightDelays$Delay[FlightDelays$Carrier=="AA"])
hist(FlightDelays$Delay[FlightDelays$Carrier=="UA"])
qqnorm(FlightDelays$Delay[FlightDelays$Carrier=="AA"])
qqnorm(FlightDelays$Delay[FlightDelays$Carrier=="UA"])
```









From the boxplot, both airlines have a significant amount of upper outliers. However the outliers from AA are further out. From the histogram, it seems that the delays follow for each of the airlines follow poisson distributions. The QQnorm plots concur that the distributions are not normal, but the two airlines follow the same distribtuion.

Part B

```
t.test(FlightDelays$Delay[FlightDelays$Carrier=="AA"], FlightDelays$Delay[FlightDelays$Carrier=="UA"])$c
## [1] -8.903198 -2.868194
## attr(,"conf.level")
## [1] 0.95
aa = FlightDelays$Delay[FlightDelays$Carrier=="AA"]
ua = FlightDelays$Delay[FlightDelays$Carrier=="UA"]
BootstrapMedianDifference = replicate(10000,
                                      median(sample(aa, length(aa),replace = T))-
                                        median(sample(ua, length(ua), replace = T)))

quantile(BootstrapMedianDifference, c(0.025,0.975))
##  2.5% 97.5%
##    -2    -1
```

The two tests seem to contradict one another because the upper bound of the t test is greater the confidence interval of the mean. This is due to the confidence interval being based on the difference in median whereas the t test is based on the difference in mean.