Homework 5

Erik Lee

Mon Jul 30 16:59:21 2018

Problem 1

A production line produces rulers that are supposed to be 12 inches long. A sample of 49 of the rulers had a mean of 12.1 and a standard deviation of .5 inches. The quality control specialist responsible for the production line decides to do a two-sided hypothesis test to determine whether the production line is really producing rulers that are 12 inches long or not.

a) What is the null hypothesis?

H_o: mu = 12 inches,

b) What is the alternative hypothesis

H_a: mu != 12 inches

c) Using whatever method you want, clearly run and summarize the result of the hypothesis test. What does this mean in terms of the problem situation?

```
#install.packages("BSDA")
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
tsum.test(n.x=49,mean.x=12,s.x=0.5,mu=12.1)
```

```
## Warning in tsum.test(n.x = 49, mean.x = 12, s.x = 0.5, mu = 12.1): argument
## 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = -1.4, df = 48, p-value = 0.2
## alternative hypothesis: true mean is not equal to 12.1
## 95 percent confidence interval:
##  11.86 12.14
## sample estimates:
## mean of x
##        12
```

Based on a two tailed t-test with a significance level of 5%, there is not enough evidence to reject the null hypothesis. The computed t=-1.4 is greater than the -1.96, supporting the null hypothesis. The p-value of 0.2 being larger than 0.05 also suggests supporting the null hypothesis based on the sample data. In terms of the problem, this data shows that it is plausible that the assembly line is producing rulers that are on average 12 inches long.

Problem 2

A particular brand of tires claims that its deluxe tire averages more than 50,000 miles before it needs to be replaced. A survey of owners of that tire design is conducted. From the 30 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9800 miles. Do the data support the claim at the 5% level? Test using whatever method you want . Be sure to clearly state your conclusion.

H_o: mu = 50,000 miles H_a: mu > 50,000 miles

```
tsum.test(n.x=30,mean.x=46500,s.x=9800,mu=50000,alt="greater")
```

```
## Warning in tsum.test(n.x = 30, mean.x = 46500, s.x = 9800, mu = 50000, alt
## = "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = -2, df = 29, p-value = 1
## alternative hypothesis: true mean is greater than 50000
## 95 percent confidence interval:
##  43460    NA
## sample estimates:
## mean of x
##     46500
```

Based on a one-tailed "greater than" hypothesis test with a 5% level of significance, the sample data suggests there is not enough evidence to support the null hypothesis and to accept the alternative hypothesis. The t-value is -2 and less than -1.64 suggest there is not enough evidence to support the null hypothesis from the sample data. As a conclusion, the sample data substantiates the claim that the deluxe tires average more than 50,000 miles before they need to be replaced.

Problem 3

A recent study stated that if a person smoked, the average of the number of cigarettes he or she smoked was 14 per day. A researcher wanted to test the claim that the mean number was actually different from 14. A random sample of 40 smokers was obtained and found that the mean number of cigarettes smoked per day was 18. The standard deviation of the sample was 6. Using whatever hypothesis testing method you want, can you conclude that the mean number of cigarettes a person smokes per day actually different from 14?

H_o: mu = 14 cigs/day H_a: mu != 14 cigs/day

The signifcance level for the two tailed hypthosis is 5%. If the absolute value of the t is greater than 1.96 or p is less than 0.05, then we reject the null hypothesis. If not, there is not enough evidence from the sample to reject the null.

```
# Calculating t by hand
# n=40, x_bar=18, s=6, mu=14
(18-14)/(6/sqrt(40))
```

```
## [1] 4.216
```

```
# Getting t-value and p-value from tsum.test() function of BSDA package
tsum.test(n.x=40,mean.x=18,s.x=6,mu=14)
```

```
## Warning in tsum.test(n.x = 40, mean.x = 18, s.x = 6, mu = 14): argument
## 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 4.2, df = 39, p-value = 1e-04
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  16.08 19.92
## sample estimates:
## mean of x
##        18
```

Based on the calculation and tsum.test() function we find that t=4.216. Further, p=1e-04 is close to 0. The t-value is greater than 1.96 and p-value is less than 0.05. These results suggest that the sample data does not support the null hypothesis and accepts the alternative hypothesis. The claim that a smoker smokes on average 14 cigarettes a day is substantiated.

Problem 4

In this exercise we show the relationship between sample size and sample evidence. Suppose the nationwide average for the math SAT test is 480 but we think UCLA students are smarter than the average. We want to test $H_o : \mu = 480$ versus $H_a : \mu > 480$.

a) Using R what is the p-value for this test if $n = 100, \hat{x} = 483, s = 100$.

```
tsum.test(n.x=100,mean.x=483,s.x=100,mu=480,alt="greater")
```

```
## Warning in tsum.test(n.x = 100, mean.x = 483, s.x = 100, mu = 480, alt =
## "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 0.3, df = 99, p-value = 0.4
## alternative hypothesis: true mean is greater than 480
## 95 percent confidence interval:
##  466.4    NA
## sample estimates:
## mean of x
##       483
```

p-value=0.4

b) Using R what is the p-value for this test if $n = 1000, \hat{x} = 483, s = 100$.

```
tsum.test(n.x=1000,mean.x=483,s.x=100,mu=480,alt="greater")
```

```
## Warning in tsum.test(n.x = 1000, mean.x = 483, s.x = 100, mu = 480, alt =
## "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 0.95, df = 1000, p-value = 0.2
## alternative hypothesis: true mean is greater than 480
```

```
## 95 percent confidence interval:
##  477.8    NA
## sample estimates:
## mean of x
##       483
```

p-value=0.2

c) Using R what is the p-value for this test if $n = 10000, \hat{x} = 483, s = 100$.

```
tsum.test(n.x=10000,mean.x=483,s.x=100,mu=480,alt="greater")
```

```
## Warning in tsum.test(n.x = 10000, mean.x = 483, s.x = 100, mu = 480, alt =
## "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 3, df = 10000, p-value = 0.001
## alternative hypothesis: true mean is greater than 480
## 95 percent confidence interval:
##  481.4    NA
## sample estimates:
## mean of x
##       483
```

p-value=0.001

d) What happens to the p-value as the sample size increases?

As sample size increases, the p-value decreases.

Problem 5

Suppose the same set up as above with the same hypothesis to test. a) Test $H_o : \mu = 480$ versus $H_a : \mu > 480$ assuming $n = 100, \hat{x} = 496.4, s = 100$.

```
tsum.test(n.x=100,mean.x=496.4,s.x=100,mu=480,alt="greater")
```

```
## Warning in tsum.test(n.x = 100, mean.x = 496.4, s.x = 100, mu = 480, alt =
## "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 1.6, df = 99, p-value = 0.05
## alternative hypothesis: true mean is greater than 480
## 95 percent confidence interval:
##  479.8    NA
## sample estimates:
## mean of x
##     496.4
```

p-value=0.05

b) Test $H_o : \mu = 480$ versus $H_a : \mu > 480$ assuming $n = 100, \hat{x} = 496.7, s = 100$.

```
tsum.test(n.x=100,mean.x=496.7,s.x=100,mu=480,alt="greater")
```

```
## Warning in tsum.test(n.x = 100, mean.x = 496.7, s.x = 100, mu = 480, alt =
## "greater"): argument 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 1.7, df = 99, p-value = 0.05
## alternative hypothesis: true mean is greater than 480
## 95 percent confidence interval:
##  480.1    NA
## sample estimates:
## mean of x
##     496.7
```

p-value=0.05

c) In practical terms should there be a difference between (a) and (b)? Discuss briefly.

Yes, in a practical sense, there should be a difference between (a) and (b) because the means, x_hat, differ by 0.3, but all other parameters (n, s, mu) are equal. (b) has a higher x_hat of 496.7 compared to (a)'s x_hat of 496.4. This factors into the t-value and p-value, when computed. But oddly, when a tsum.test() function is run for both, the p-value for both is 0.05 likely due to rounding. This is a problem because 0.05 is considered the cut off point for a p-value to determine if the null hypothesis is supported or not supported.

If we do the hand calculation with R, we can see exactly how these p-values differ:

```
# for (a)
z_a = (496.4-480)/(100/sqrt(100))
1-pnorm(z_a) # right-side of the normal distribution, subtract the area from 1
```

```
## [1] 0.0505
```

```r
# for (b)
z_b = (496.7-480)/(100/sqrt(100))
1-pnorm(z_b)
```

```
## [1] 0.04746
```

We see from the calculations, the p-value of (a) and (b) are different. Using the p=0.05 rule, we would support the null hypothesis for (a) and not support the null hypothesis for (b). This highlights a problem with the p=0.05 rule and sample data. The 0.05 rule seems arbitrary because the values for (a) and (b) are essentially equal (n, mean, s, mu) with only x_hat differing by 0.3. It may be better to judge how close the p-value is to 0 to get a better sense of whether to not support the null hypothesis. Or we could use a more descriptive statstical test for the hypothesis.

Problem 6

A survey of 4000 people in the US finds that 2856 of them believe that daily weather reports are totally useless because meteorology is not really a science. Given this data perform a hypothesis test to see if more than half of the people in the US believe that weather reports are useless.

d) What is the null hypothesis?

H_o: p_true = 0.5

e) What is the alternative hypothesis

H_a: p_true > 0.5

f) Using whatever method you want, clearly run and summarize the result of the hypothesis test. What does this mean in terms of the problem situation?

To test this claim, the t-value will be calculated. Since this is a one tailed hypothesis test, where the claim is that the proportion is greater than 0.5, the null hypothesis will be rejected if t_stat is greater than 1.64.

```r
# p_hat=2856/4000=0.714, p_o=0.5
(0.714-0.5)/(sqrt(0.5*(1-0.5)/4000))
```

```
## [1] 27.07
```

```r
prop.test(x=2856,n=4000,p=0.5,alt="greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  2856 out of 4000, null probability 0.5
## X-squared = 730, df = 1, p-value <2e-16
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.702 1.000
## sample estimates:
##     p
## 0.714
```

t=27.07 p-value<2e-16

Based on the test, t=27.07 was returned. With this value being greater than 1.64, we say the sample data does not support the null hypothesis and accept the alternative hypothesis. With a prop.test() function, p-value<2e-16, p-value is almost 0. This too is evidence that the data does not support the null hypothesis.

Based on a 5% statistical significance level, the claim that half the people in US believe the weather reports are useless is plausible based on the sample data provided by the survey of 4000 people in the US.

Problem 7

A poll done for Newsweek found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only 2 had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the Newsweek poll?

H_o: p_true = 0.13 H_a: p_true != 0.13

```
# p_hat=2/76=0.02631, p=0.13, n=76
((2/76)-0.13)/(sqrt(0.13*(1-0.13)/76))
```

```
## [1] -2.688
```

```
prop.test(x=2,n=76,p=0.13)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  2 out of 76, null probability 0.13
## X-squared = 6.3, df = 1, p-value = 0.01
## alternative hypothesis: true p is not equal to 0.13
## 95 percent confidence interval:
##  0.004571 0.100488
## sample estimates:
##       p
## 0.02632
```

Based on a two sided hypothesis, t=-2.688 and p=0.01. Since the absolute value of t is greater than 1.96 and p-value is less than 0.05, the null hypothesis is not supported and the alternative hypothesis can be accepted. Based on a 5% significance level, the claim that 13% of Americans have seen or sense the presence of an angel is not plausible based on the data from the survey of 76 Americans.

Problem 8

According to the 2010 Census, 58.5% of women worked. A county commissioner feels that more women work in his county, so he conducts a survey of 1000 randomly selected women and finds that 622 work. Is he correct?

H_o: p_true = 0.585 H_a: p_true > 0.585

```
# p_hat=622/1000=0.622, p_o=0.585
(0.622-0.585)/(sqrt(0.585*(1-0.585)/1000))
```

```
## [1] 2.375
```

```
prop.test(x=622,n=1000,p=0.585,alt="greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  622 out of 1000, null probability 0.585
```

```
## X-squared = 5.5, df = 1, p-value = 0.01
## alternative hypothesis: true p is greater than 0.585
## 95 percent confidence interval:
##  0.596 1.000
## sample estimates:
##     p
## 0.622
```

His claim that the percentage of working women is plausible based on the random survey's data. The one-tailed "greater than" hypothesis tests of the data shows that t=2.375 and p=0.01. A t-value over 1.64 and p-value less than 0.05 would not support the null hypothesis. With a significance level of 5%, the claim can be supported that the percentage of women who worked in 2010 is higher than the 58.5% reported by the Census.

Problem 9

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

H_o: p_true = 0.40 H_a: p_true < 0.40

```
((135/361)-0.40)/(sqrt(0.40*(1-0.40)/361))
```

```
## [1] -1.01
```

```
prop.test(x=135,n=361,p=0.40,alt="less")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  135 out of 361, null probability 0.4
## X-squared = 0.91, df = 1, p-value = 0.2
## alternative hypothesis: true p is less than 0.4
## 95 percent confidence interval:
##  0.000 0.418
## sample estimates:
##     p
## 0.374
```

Based on a 5% level of significance and a one-tailed "less than" t-test, t=-1.01 and p=0.2. Since the absolute value of t=-1.01 is less than 1.64 and p=0.2 is greater than 0.05, the test suggests that there is evedonce to support null hypthesis and reject the alternative hypothesis that less than 40% of the students in the surveyor's school fears public speaking.

Problem 10

Two groups of students are given a problem-solving test, and the results are compared. Find and interpret the 95% confidence interval of the true difference in means. Feel free to use R.

Mathematics majors	Computer science majors
$\bar{X}_1 = 83.6$	$\bar{X}_2 = 79.2$
$s_1 = 4.3$	$s_2 = 3.8$
$n_1 = 36$	$n_2 = 36$

H_o: x_mu = y_mu H_a: x_mu != y_mu

x_mu = true mean for math majors y_mu = true mean for cs majors

```
tsum.test(n.x=36,mean.x=83.6,s.x=4.3,n.y=36,mean.y=79.2,s.y=3.8)
```

```
##
##  Welch Modified Two-Sample t-Test
##
## data:  Summarized x and y
## t = 4.6, df = 69, p-value = 2e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.492 6.308
## sample estimates:
## mean of x mean of y
##      83.6      79.2
```

Using the tsum.test() function, the 95% confidence interval for the true difference in mean is [2.492,6.308]. This interval represents the potential value of x_mu-y_mu, where x_mu is the true mean for math majors and y_mu is the true mean for cs majors. Since the difference is positive, data supports that math majors have a greater true mean than the cs major true mean.

This is a two sided hypothesis, where the null hypothesis is that the true means for each respective group is equal. The alternative hypothesis is that the true means are not equal. From the tsum.test() function we are given a t=4.6 and p-value=2e-05. Since t is greater than 1.96 and p is less than 0.05, these results suggest not supporting the null hypothesis and accepting the alternative hypothesis. This test say it is plausible the true means for math majors and cs majors are not equal, or the true difference between the means is not 0.

Problem 11

Many doctors believe that early prenatal care is very important to the health of a baby and its mother. Efforts have recently been focused on teen mothers. A random sample of 52 teenagers who gave birth revealed that 32 of them began prenatal care in the first trimester of their pregnancy. A random sample of 209 women in their twenties who gave birth revealed that 163 of them began prenatal care in the first trimester of their pregnancy.

a. Construct a 95% confidence interval for the difference between the proportion of teen mothers who get early prenatal care and the proportion of mothers in their twenties who get early prenatal care. (you may do this using R).

H_o: p_1 = p_2 H_a: p_1 != p_2

p_1 = true proportion of twenties women using early prenatal care P_2 = true proportion of teenage women using early prenatal care

```
prop.test(c(32,163),c(52,209))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(32, 163) out of c(52, 209)
## X-squared = 5.1, df = 1, p-value = 0.02
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.320194 -0.008846
## sample estimates:
## prop 1 prop 2
## 0.6154 0.7799
```

```
# calculate T
p_hat=((52*0.6154)+(209*0.7799))/(52+209)
(0.6154-0.7799)/sqrt(p_hat*(1-p_hat)*((1/52)+(1/209)))
```

```
## [1] -2.442
```

b. Briefly interpret the confidence interval. Using the prop.test() for a 2-sample test for equality of proportion, the 95% confidence interval for the difference is [-0.320194,-0.008846]. This interval represents a 95% confidence that the value of true difference in proportion is found within this range. This interval has a negative range and describes that the difference between p_1 for teenage pregnancies and p_2 for 20's age pregnancies is negative.

This negative difference in proportion indicates that p_2 proportion for twenties age pregnancies is larger than p_1 proportion for teenage pregnancies. In other words, the sample data indicates that it is plausible a larger true proportion pregnant women in their 20's applied early prenatal care to their babies compared to the true proportion for teenage pregnancies.

Problem 12

In a sample of 80 Americans, 55% wished that they were rich. In a sample of 90 Europeans, 45% wished that they were rich. Run a two sided hypothesis test to see if there is a difference in the proportions against the null that they are equal.

H_o: p_1 = p_2 H_a: p_1 != p_2

p_1 = true proportion of Americans wishing they were rich p_2 = true proportion of Europoeans wishing they were rich

```
prop.test(c(44,40.5),c(80,90))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(44, 40.5) out of c(80, 90)
## X-squared = 1.3, df = 1, p-value = 0.3
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06163  0.26163
```

```
## sample estimates:
## prop 1 prop 2
##   0.55   0.45
```

```
# calculate T
p_hat=((80*0.55)+(90*0.45))/(80+90)
(0.55-0.45)/sqrt(p_hat*(1-p_hat)*((1/80)+(1/90)))
```

```
## [1] 1.302
```

Using a two tailed hypothesis test, the difference of true proportions are calculated in terms of a t-value and p-value. The p-value obtained by the prop.test() function is 0.3 and the t-value, using the equation to calculate T, is 1.302. Since the p-value is greater than 0.05 and t is less than 1.96, the sample data supports the null hypothesis. So the data concludes wtih a 5% level of significance that it is plausible the true proportion of Americans wishing they were rich and the true proportion of Europeans wishing they were rich could be equal (p_1 = p_2). This is also supported by the 95% confidence interval spanning the 0 value, meaning the true difference in proportions can equal 0.

Problem 13

Suppose in a survey of college students, 1630 out of 7180 men responded Yes to being frequent binge drinkers and 1684 out of 9916 women responded yes. Find a 95% confidence interval for the difference between the proportions of men and women who are frequent binge drinkers. Interpret the interval.

H_o: p_m = p_w H_a: p_m != p_w

p_m: true proportion for college men who binge drink p_w: true proportion for college women who binge drink

```
prop.test(c(1640,1684),c(7180,9916))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(1640, 1684) out of c(7180, 9916)
## X-squared = 91, df = 1, p-value <2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04626 0.07091
## sample estimates:
## prop 1 prop 2
## 0.2284 0.1698
```

The 95% confidence interval for the difference in men's and women's proportion of binge drinking is [0.04626,0.07091]. The range of confidence interval is positive. Since this interval represents the difference in proportion of p_m-p_w and the interval is positive, the survey data suggests with a 5% significance level, that there is a larger true proportion of men who binge drink compared to the true proportion of women who binge drink at this college.

Problem 14

In an effort to increase production of an automobile part, the factory manager decides to play music in the manufacturing area. Eight workers are selected, and the number of items each produced for a specific day is recorded. After one week of music, the same workers are monitored again. The data are given below. Can the manager conclude that the music has increased?

Worker	1	2	3	4	5	6	7	8
Before	6	8	10	9	5	12	9	7
After	10	12	9	12	8	13	8	10

mu_bf: true mean for number of items produced by workers before music is played mu_at: true mean for number of items produced by workers after music is played

H_o: mu_bf = mu_at H_a: mu_bf < mu_at

```
workers_before = c(6,8,10,9,5,12,9,7)
workers_after = c(10,12,9,12,8,13,8,10)
t.test(workers_after,workers_before,alt="greater",paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  workers_after and workers_before
## t = 2.7, df = 7, p-value = 0.01
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6133    Inf
## sample estimates:
## mean of the differences
##                       2
```

Based on a one-sided "greater than" paired hypothesis test with a significance level of 5%, the conclusion that playing music in the manufacturing area increases average item production is plausible based on the sample data collected by the factory manager. Using the t.test() function, t=2.7 and p=0.01. Since t is greater than 1.64 and p is less than 0.05, the null hypothesis is not supported and the alternative hypothesis is accepted. The manager is able to conclude, based on the data, that music improves the average production of items in the factory.

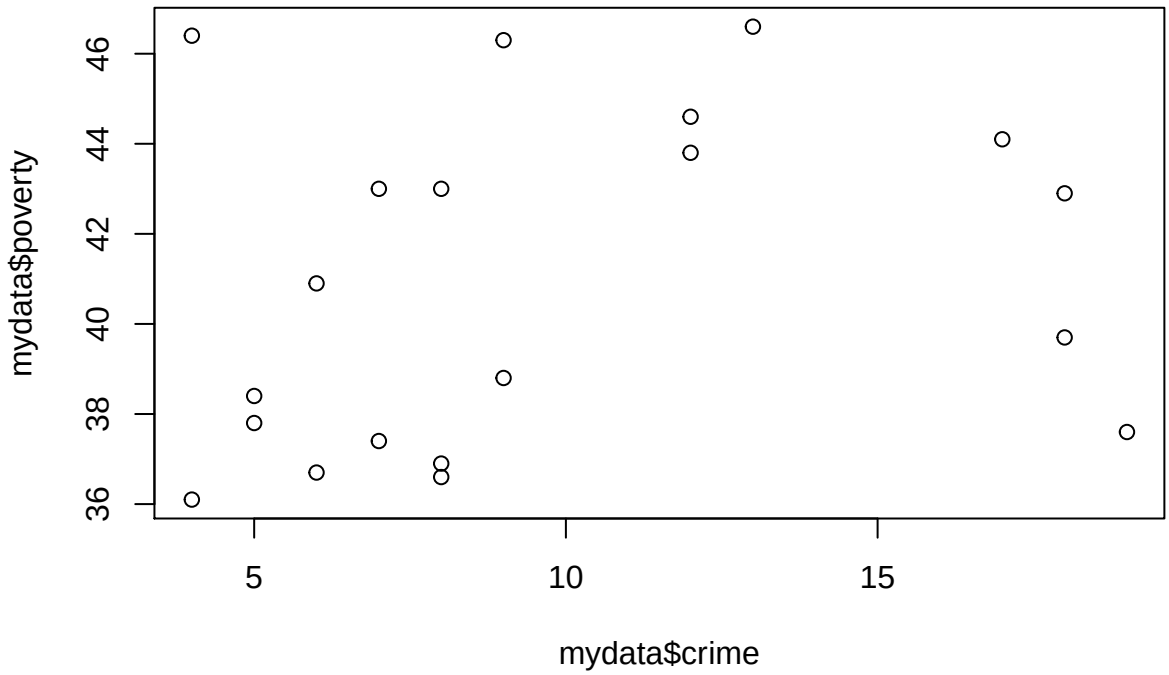
Problem 15

A sociologist is curious if the poverty level of children in large cities can be predicted with a linear model based on the city's crime rate. Load the data into R using the command

mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/poverty.csv")

a) Using Poverty as the "Y" variable and Crime as the "X" variable, create a scatter plot. Does there appear to be a linear relationship?

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/poverty.csv")
plot(x=mydata$crime,y=mydata$poverty)
```



It does not seem that there is clear linear relationship from the scatter plot. The points are spread out well and there is a collection of low crime and low poverty, low crime and high poverty, high crime and low poverty, and high crime and high poverty points across the plot.

b) What is the value of R-sq? Would you say it is a high or low value?

```
regress_line=lm(mydata$poverty~mydata$crime)
summary(regress_line)
```

```
##
## Call:
## lm(formula = mydata$poverty ~ mydata$crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.954 -3.112 -0.547  2.539  6.561
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    39.116      1.858   21.05    4e-14 ***
## mydata$crime    0.181      0.171    1.06      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 18 degrees of freedom
## Multiple R-squared:  0.0586, Adjusted R-squared:  0.00627
## F-statistic: 1.12 on 1 and 18 DF,  p-value: 0.304
```

R-sq=0.0586 and adjusted R-sq=0.00627. This is a low value for R-sq and adjusted R-sq.

c) Test the hypothesis . Fully explain your conclusion.

```
regress_line=lm(mydata$poverty~mydata$crime)
summary(regress_line)
```

```
##
```

```
## Call:
## lm(formula = mydata$poverty ~ mydata$crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.954 -3.112 -0.547  2.539  6.561
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    39.116      1.858   21.05    4e-14 ***
## mydata$crime    0.181      0.171    1.06      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 18 degrees of freedom
## Multiple R-squared:  0.0586, Adjusted R-squared:  0.00627
## F-statistic: 1.12 on 1 and 18 DF,  p-value: 0.304
```

H_o: beta_1 = 0, no linear relationship between crime and poverty H_a: beta_1 != 0, possible linear relationship between crime and poverty

beta_1 = true value for the slope of the regression line

The hypothesis is if poverty level of children can be predicted with the crime rate of a city, based on a linear regression model. Using the lm() function in R to create a linear model and summarizing the results, it is not plausible to support the hypothesis that crime rate can predict poverty level of children with a linear model.

There are several pieces of evidence to support this claim. The R-sq value is 0.0586 a very low value, and while R-sq is not the best a predicting fit of a line, it does indicate a weakness in the linear model. The t-value for mydata$crime (b_1) is 1.06. This is less than 1.96. And the p-value for mydata$crime is 0.3, greater than 0.05. Both the t-value and p-value indicate that the null hypothesis is supported by data and it is plausible that there is not relationship between poverty and crime.

d) Should we use this regression equation to predict children's poverty levels from the crime rates? Explain.

Based on the explanation of part C, this is not a very good linear model to predict children's poverty from crime rate. We saw above based on R-sq, t-value, and p-value for mydata$crime (b_1) there is not enough statistical evidence to reject the null hypothesis.

```
regress_line=lm(mydata$poverty~mydata$crime)
confint(regress_line, level=0.95)
```

```
##                 2.5 %  97.5 %
## (Intercept)  35.2119 43.0192
## mydata$crime -0.1783  0.5402
```

This is further supported by creating a 95% confidence interval for mydata$crime (b_1) [-0.1783,0.5402]. This interval spans 0, with a negative min and positive max. This allows for the possibility that b_1=0, in which case there is no slope and no relationship between crime and poverty. So this linear model is not adequate at explaining a relationship for crime and poverty of children.

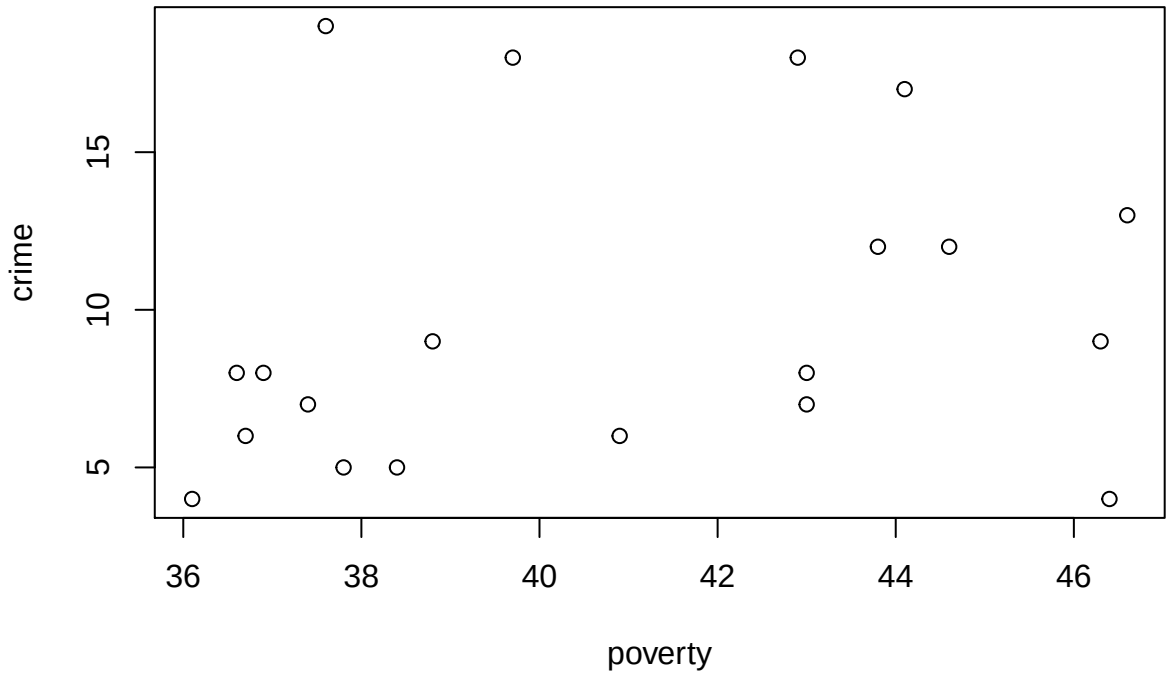
Problem 16

A linear model was proposed to relate the top offensive linemen for the NFL draft to their times in the 40-year dash. Load the data into R using the command

mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/nfldata.csv")

a) Create a scatter plot. Does there appear to be a linear relationship?

```
plot(mydata)
```



There does not seem to be a clear linear relationship for rating and time, based on the plot. A low rating point seems to have the same time as some high rating points. It is difficult to tell if the time can be predicted by rating.

b) What is the value of R-sq? Would you say it is a high or low value?

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/nfldata.csv")
fit=lm(mydata$time~mydata$rating)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$time ~ mydata$rating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2653 -0.0393  0.0152  0.0453  0.3055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.5250     0.2077   26.60   <2e-16 ***
## mydata$rating  -0.0492     0.0340   -1.45     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.118 on 27 degrees of freedom
## Multiple R-squared:  0.0721, Adjusted R-squared:  0.0377
## F-statistic:  2.1 on 1 and 27 DF,  p-value: 0.159
```

R-sq = 0.0721. This is a low R-sq value.

c) Test the hypothesis $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.. Fully explain your conclusion.

Based on the linear regression model with a 5% level of signficiance, there is not enough evidence to reject the null hypothesis using the sample data. The R-sq value is 0.0721 indicating a poor linear model for predicting the relationship between player rating and 40-yard dash time. The rating has a t-value of -1.45, which has an aboslute value lower than 1.96. And the p-value for rating is 0.16, which is greater than 0.05. Both the t-value and p-value show that rejection of the null hypothesis can not be supported with the given data.

d) Does this appear to be a useful model? Explain.

Based on the conclusion from part C, this model is not useful for predicting the value of 40-yard dash times based on player rating. The low R-sq value and adjusted R-sq of 0.0377 suggest the model is not strong for predicting the linear relationship of the data. The t-value and p-value highlight that there is not enough evidence to reject the null hypothesis.

```
fit=lm(mydata$time~mydata$rating)
confint(fit, level=0.95)
```

```
##                  2.5 %  97.5 %
## (Intercept)    5.0988 5.95110
## mydata$rating -0.1188 0.02049
```

If we look at the confidence interval of the data, we see that the mydata$rating (b_1) interval spans 0. This means the interval includes the 0 value and b_1 may assume this value, indicating no relationship between X (player rating) and Y (40-yard dash times). Since there is a potential of no linear relationship, this regression proves to be a poor prediction model.

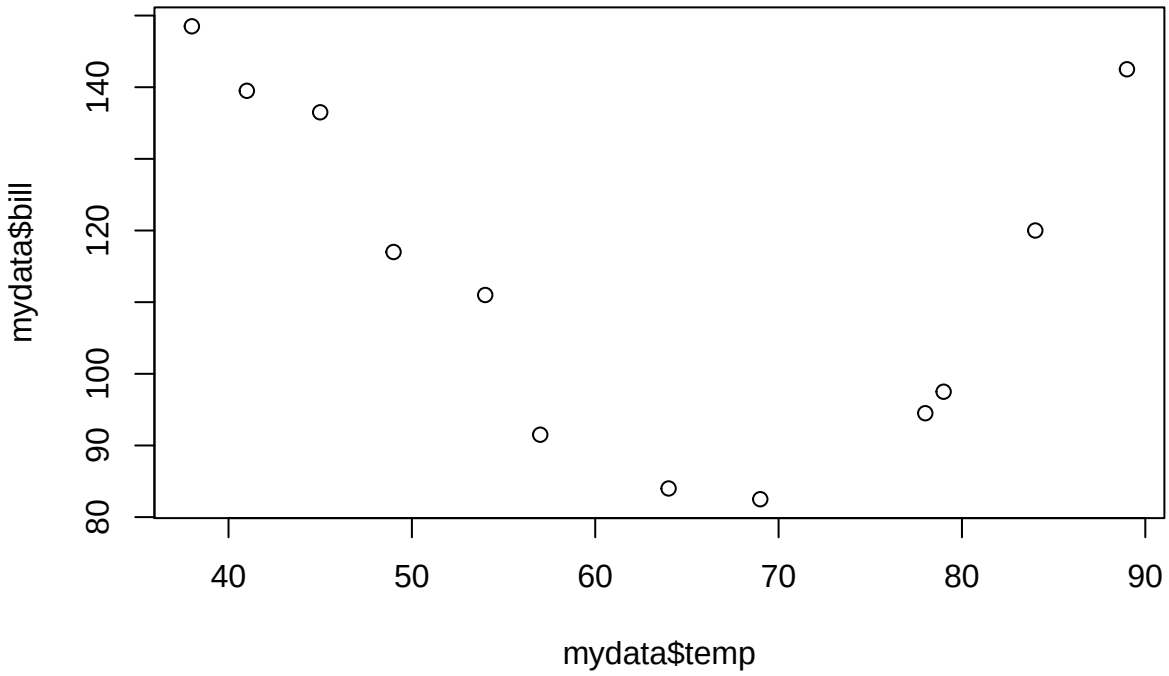
Problem 17

In data file utility.xls are the average utility bills for homes of a particular size (Y) and the average monthly temperature (X). Use R to answer the questions below. Load the data into R using the command

mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/utility.csv")

a) Make a scatter plot of the data.

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/utility.csv")
plot(x=mydata$temp, y=mydata$bill)
```



b) Does it appear from inspection that there is a relationship between the variables? Why or why not?

There seems to be a relationship between temp and bill, but not a linear relationship. The graph curves creating a parabola, so it may be the case that temp and bill have a quadratic relationship.

c) Using R, calculate the least squares line.

```
fit=lm(mydata$bill~mydata$temp)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$bill ~ mydata$temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28.91 -14.96  -4.91  15.84  41.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  143.623     25.995    5.52  0.00025 ***
## mydata$temp   -0.480      0.403   -1.19  0.26149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 23.4 on 10 degrees of freedom
## Multiple R-squared:  0.124,  Adjusted R-squared:  0.0365
## F-statistic: 1.42 on 1 and 10 DF,  p-value: 0.261
```

Least Squares Equation: Y = 143.623 + (-0.480)*X

b_0 = 143.623, b_1 = -0.480 X = average monthly temperature, Y = average utility bills

d) Using R, calculate and interpret the value of R^2.

```
fit=lm(mydata$bill~mydata$temp)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$bill ~ mydata$temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28.91 -14.96  -4.91  15.84  41.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  143.623     25.995    5.52  0.00025 ***
## mydata$temp   -0.480      0.403   -1.19  0.26149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.4 on 10 degrees of freedom
## Multiple R-squared:  0.124,  Adjusted R-squared:  0.0365
## F-statistic: 1.42 on 1 and 10 DF,  p-value: 0.261
```

R-sq = 0.124. Adjusted R-sq = 0.0365. Both these low values indicating a potential weakness in the predictive ability of this linear model for the given sample data.

e) Test the hypothesis $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Fully explain your conclusion.

Based on a two tailed hypothesis test with a 5% level of significance for the data, there is not enough evidence to reject the null hypthesis that beta_1=0. The t-value for b_1 is -1.19, which has an absolute value of 1.19. This absolute t-value is less than 1.96 so we can not reject the null hypothesis. The p-value for b_1 is 0.26149, which is greater than 0.05. This evaluation also says we do not have evidence to reject the null hypothesis. These two values indicate that the null hypothesis is plausible and that the relationship between average montly temp and average utility costs could have a 0 relationship.

f) Use the least squares line to estimate the average utility bill if the average monthly temperature is 120 degrees. Do you think that your answer is reasonable? Why or why not?

```
143.623 + (-0.480)*(120)
```

```
## [1] 86.02
```

The bill for an average monthly temp of 120 is $86.02. Based on the plot of the data, this does not seem to be a reasonable answer. If we mapped an average bill of $86.02 to the scatter plot, the approximate temperature is around 65-70 degrees, almost 50 degrees lower than 120.

Also, if we look at how the graph progresses from 80 to 90 degrees for temperature, the bill value (Y) seems to be increasing. So if the graph was extended further to include 120 degree as a temp point, the bill would increase further above $140 for the bill.

This graph makes some sense as average room temperature is 73 degrees Fahrenheit. If temperature is much lower, heat would be turned on. If temperature is much higher, air conditioning will be turned on. Both heating and air conditioning drain electricity and increase monthy bill cost.

g) Give a 95% prediction interval for the utility bill if the temperature is 60 degrees.

```
# Y = 143.623 + (-0.480)*(X), X = 60
# 95% predict. interval = Y +/- 1.96*s_e
# s_e = residual standard error = 23.4
# Y value
Y = 143.623 + (-0.480)*(60)
# min
Y-1.96*(23.4)
```

```
## [1] 68.96
```

```
# max
Y+1.96*(23.4)
```

```
## [1] 160.7
```

95% prediction interval for utility bill at 60 degrees temp = [68.96,160.7]

Problem 18

In simple linear regression, what is the difference between b_1 and β_1 ?

b_1 is the estimated value for beta_1. beta_1 is the true value for the slope of a simple linear regression model.

Problem 19

A large class of 360 students has just taken an exam. The exam consisted of 40 true-false questions each of which was worth one point. A diligent teaching assistant has recorded the number of correct answers (Y) and the number of incorrect answers (X) for each student. Suppose that the student then regresses the variable Y on the variable X. What will be the values of b_0, b_1 and R^2 ? Is this a sensible model to fit to the data ? [there is no data for this problem-it isn't needed].

Y = 40 - X b_0=40, b_1=-1, Rˆ2=100%

Problem 20

We have a data file that contains data from the 2005 World Factbook relating to gross domestic product (GDP) per capita in US$ thousands (gdp) and the percentage of the population that are internet users (intpct) for 213 countries. Here, GDP is based on purchasing power parities to account for between-country differences in price levels. This problem investigates whether there is a linear association between these two variables. In particular, how effective is it to use gdp to predict intpct using simple linear regression.

You can read the data into R as follows:

mydata=read.csv("http://www.datadescant.com/stat104/internet.csv")

a) Using R, find the least squares line for the data.

```
mydata=read.csv("http://www.datadescant.com/stat104/internet.csv")
fit=lm(mydata$intpct~mydata$gdp)
summary(fit)
```

```
##
## Call:
## lm(formula = mydata$intpct ~ mydata$gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.74 -11.91  -3.28   9.42  63.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3628     1.7183    7.19  1.1e-11 ***
## mydata$gdp    1.3609     0.0797   17.07  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.1 on 210 degrees of freedom
## Multiple R-squared:  0.581,  Adjusted R-squared:  0.579
## F-statistic:  291 on 1 and 210 DF,  p-value: <2e-16
```

Y = 12.3628 + 1.3609*X b_0 = 12.3628, b_1 = 1.3609 Y = intpct, X = gdp

b) Interpret the estimates of the slope and the y-intercept in the context of the problem.

The y-intercept is 12.3628 (b_0) represents a 12.3628% of the population using internet, when gdp = 0. The slope represents the change as gdp increases by 1, the percentage of the population using the internet increases by 1.3609% (b_1).

c) Predict the percentage of internet users if GDP per capita is US$20,000.

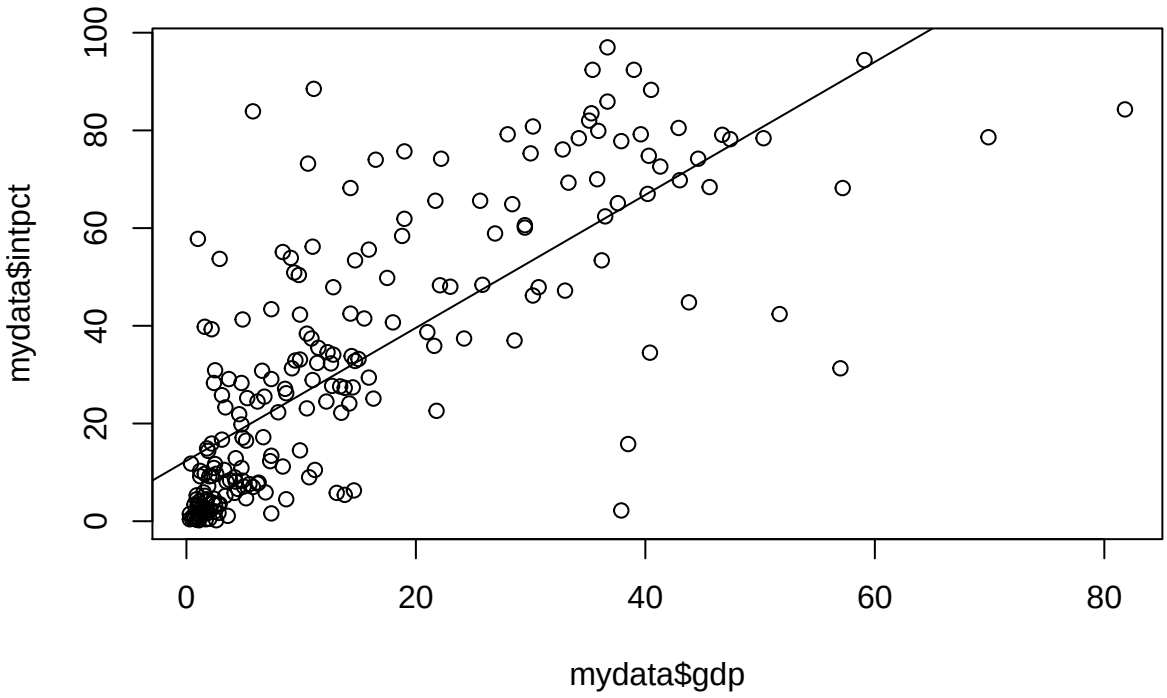
```
# $20,000 gdp = X = 20.0
12.3628 + 1.3609*(20.0)
```

```
## [1] 39.58
```

The percentage of internet users is 39.58%, if the GDP per capita is US$20,000.

d) Draw a scatterplot with intpct on the vertical axis and gdp on the horizontal axis, and add the least squares line to the plot.

```
plot(x=mydata$gdp, y=mydata$intpct)
ls_line=lm(mydata$intpct~mydata$gdp)
abline(ls_line)
```



e) Based on the scatterplot, do you think it is appropriate to use this simple linear regression model in this problem or is the model potentially misleading (and if so, how)?

The model does seem misleading. Very few points actually land on the regression line. This may be concerning as you can predict a value with the least squares equation, but the points would have a large error. The model seems to be compensating for data points that have huge errors, such that points below the line will have errors that rival the errors of points above the line. Whatever value is predicted by this regression model is likely to have a huge error and not represent true observed value. There is a lot of noise, with data points spread out in the data, so the accuracy of the regression model is low. Even though a value is returned, it would not be a reasonable prediction for a real value.

Problem 21

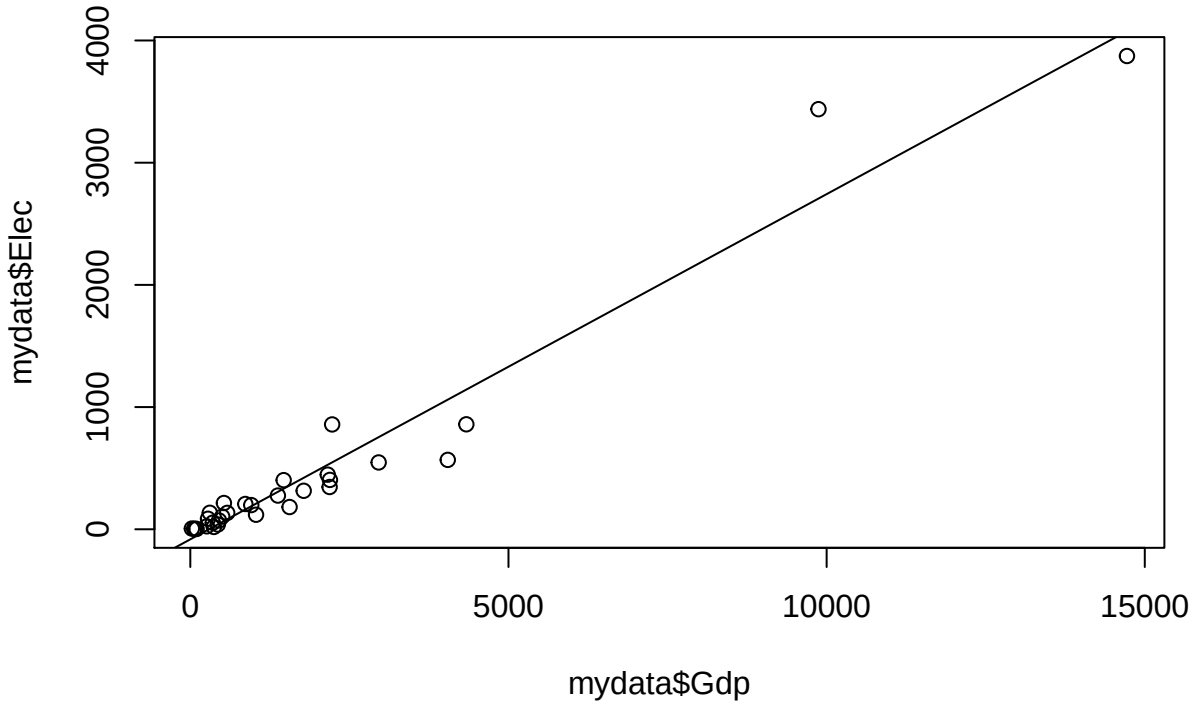
We have data on electricity consumption in billions of kilowatt-hours (Elec) and gross domestic product (GDP) in billions of dollars (Gdp) for the 30 most populous countries. Here, GDP is based on purchasing power parities to account for between-country differences in price levels. The data file can be used to investigate the claim that there is a straight-line relationship between electricity consumption and GDP. For the purposes of this problem, assume that increases in electricity consumption (Y) tend to respond to increases in GDP (X) (rather than the other way around).

You can read the data into R as follows:

mydata=read.csv("http://www.datadescant.com/stat104/electricity.csv")

a) Plot the data in a scatterplot (make sure you put the appropriate variables on each axis. Add the least squares line to the scatterplot.

```
mydata=read.csv("http://www.datadescant.com/stat104/electricity.csv")
plot(x=mydata$Gdp,y=mydata$Elec)
ls_line=lm(mydata$Elec~mydata$Gdp)
abline(ls_line)
```



b) What is the dominant pattern in the points on the scatterplot?

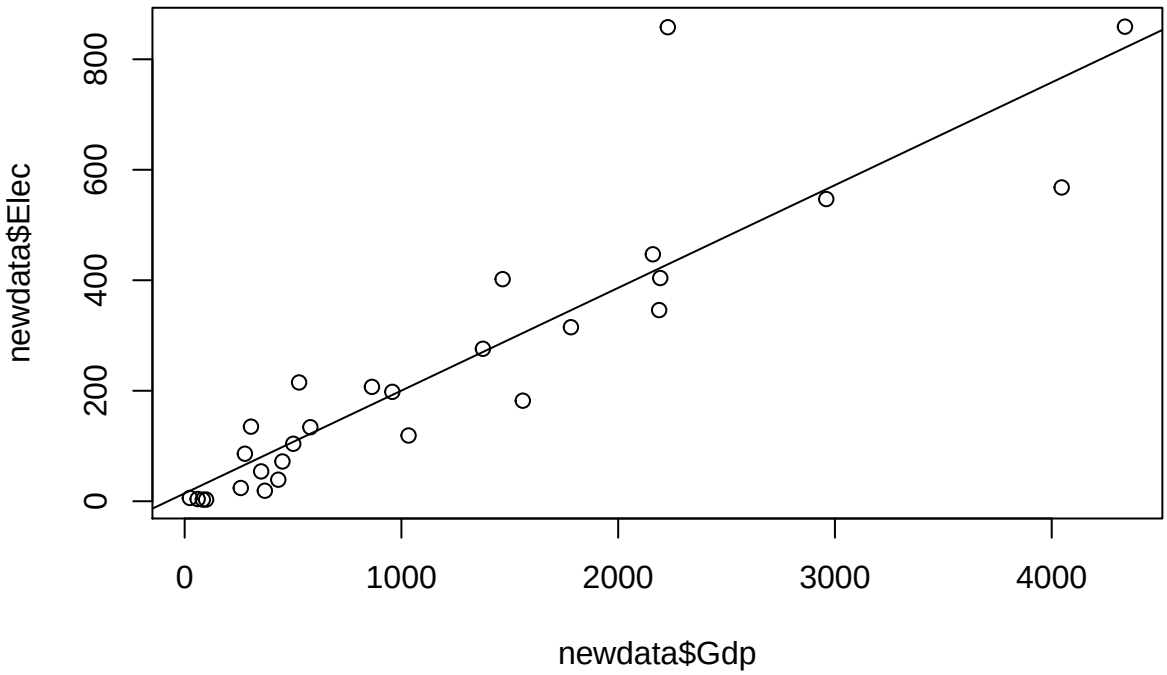
The dominant patterns is that nations with lower GDP have lower energy consumption and nations with higher GDP have higher energy consumption.

c) Identify the countries with the two highest values of GDP, remove them from the dataset, and redraw the scatterplot.

```
# finding the two countries with highest GDP
highest_gdps=subset(mydata,mydata$Gdp>8000)
print(highest_gdps)
```

```
##                X Elec   Gdp
## 4          China 3438  9872
## 29 United States 3873 14720
```

```
# removing two max values
newdata=subset(mydata,mydata$Gdp<8000)
# plot
plot(x=newdata$Gdp,y=newdata$Elec)
ls_line=lm(newdata$Elec~newdata$Gdp)
abline(ls_line)
```



Highest GDPs: China (9872 Gdp) United States (14720 Gdp)

Pro tip: One can create a new version of the data set to work with as follows:

```
> mydata=read.csv("http://www.datadescant.com/stat104/electricity.csv")
> newdata=subset(mydata,mydata$Gdp<8000)
```

d) How does your visual impression of the scatterplot change?

The data points are more spread out across the plot and least squares regression line. It is much easier to see the individual data points and there is no large gap/space between the data points, as the case with the max values and the rest of the data. They still follow the pattern mentioned above and fit the line pretty well. There seems to be a linear relationship between GDP and electricity consumption, judging by the regression line and mapping of the points.

e) Fit a least squares regression line to the new data set with the two highest values of GDP removed. Is there a significant relationship between Elec and Gdp?

```
ls_line=lm(mydata$Elec~mydata$Gdp)
summary(ls_line)
```

```
##
## Call:
## lm(formula = mydata$Elec ~ mydata$Gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -492.7 -102.2   30.0   67.4  731.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -82.2655    45.3535   -1.81     0.08 .
## mydata$Gdp    0.2825     0.0126   22.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 210 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.946
## F-statistic:  506 on 1 and 28 DF,  p-value: <2e-16
```

H_o: beta_1 = 0 H_a: beta_1 != 0

beta_1 = the true value for mydata$Gdp

The least squares regression line was included in the plot for part D. Based on analysis of the regression line, the data points fit the line well and follow the linear model for GDP and electricity consumption.

If we analyzed the summary data from the regression line, we can see a few important statistics supporting the the model. The R-sq is 0.751 and ajusted R-sq is 0.741. Both these values are high and indicate strength in this linear model. The t-value for Gdp (b_1) is 8.68. This is significantly higher than 1.96; this result is evidence to not support the null hypothesis that beta_1 = 0. In other words, there is a linear relationship between Gdp and electricty consumption. This fact is also supported by a low p-value for Gdp of 5.1e-09. This value, being close to 0, indicates a linear relationship over no relationship. Based on the statistics of a two tailed hypothesis test and 5% significance level, we can see that there is a significant linear relationship between Gdp and electricity consumption.