Problem Set 1

Ian Sims

January 14, 2019

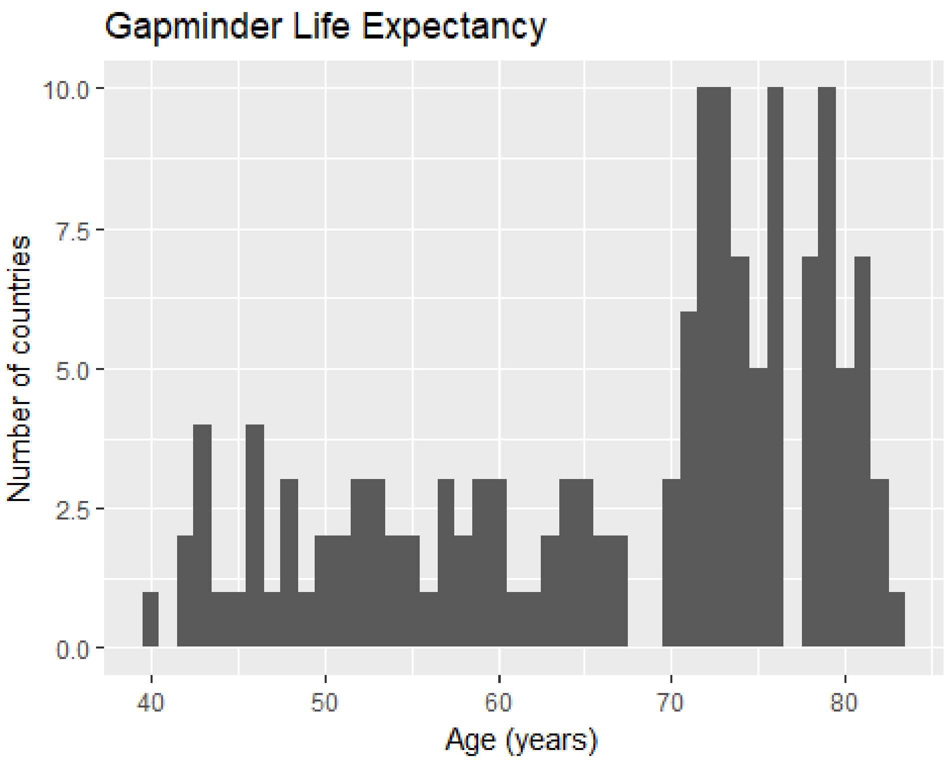
Load packages:

```
library(ggplot2)
library(lattice)
library(gapminder)
library(GGally)
library(NHANES)
library(broom)
```

Problem #1

(a) Use ggplot() to draw ONE graph of the distribution of life expectancy. Using words and numbers, describe the center, spread, and shape of the distribution.

```
gapminder07 = subset(gapminder, year == 2007)

gg = ggplot(gapminder07, aes(x = lifeExp))
gg = gg + geom_histogram(binwidth = 1)
gg = gg + labs(title = "Gapminder Life Expectancy")
gg = gg + xlab("Age (years)") + ylab("Number of countries")
gg
```

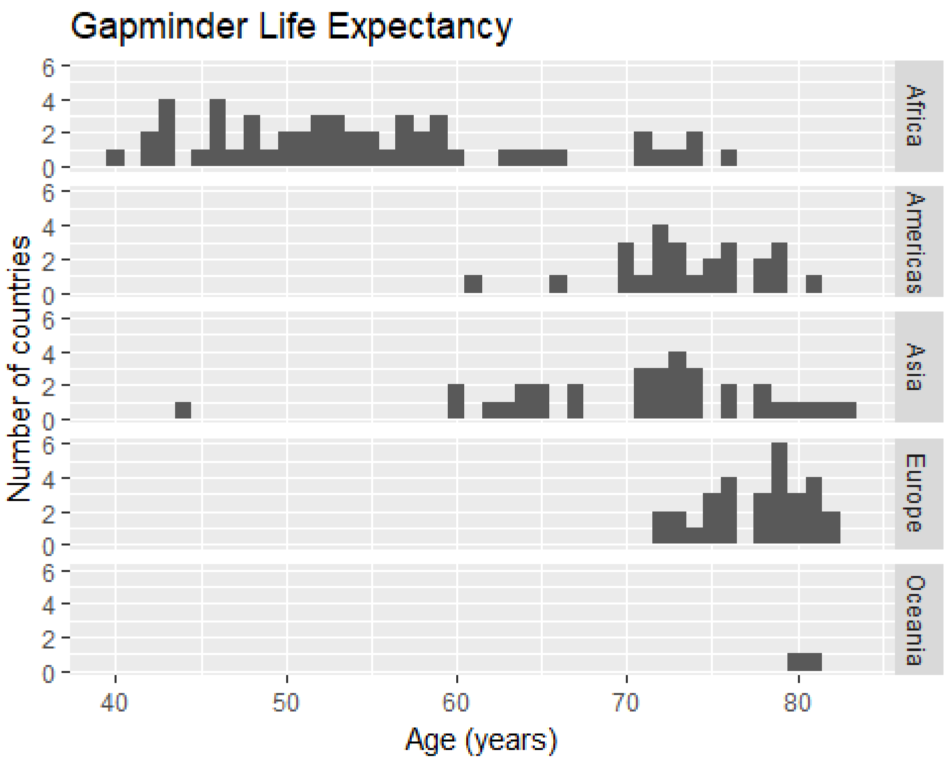


There does not seem to be a clear ‘center’ to this distribution. There seem to be more distinct distributions within the greater distribution that may have centers. For example, above 70 there appear to be two distinct clusters. One with a center around 72 and one with a center around 80.

The spread is between around 40 and 85, with more concentration between 70 and 85. here doesn’t seem to be a very disticnt shape though it is definitely right skewed.

(b) Use a set of faceted plots to display the distribution of life expectancy for each continent in a way that allows easy comparison. Are there clear differences between the distributions for different continents, or are they about the same?

```
gapminder07 = subset(gapminder, year == 2007)

gg = ggplot(gapminder07, aes(x = lifeExp))
gg = gg + geom_histogram(binwidth = 1)
gg = gg + facet_grid(continent ~ .)
gg = gg + labs(title = "Gapminder Life Expectancy")
gg = gg + xlab("Age (years)") + ylab("Number of countries")
gg
```

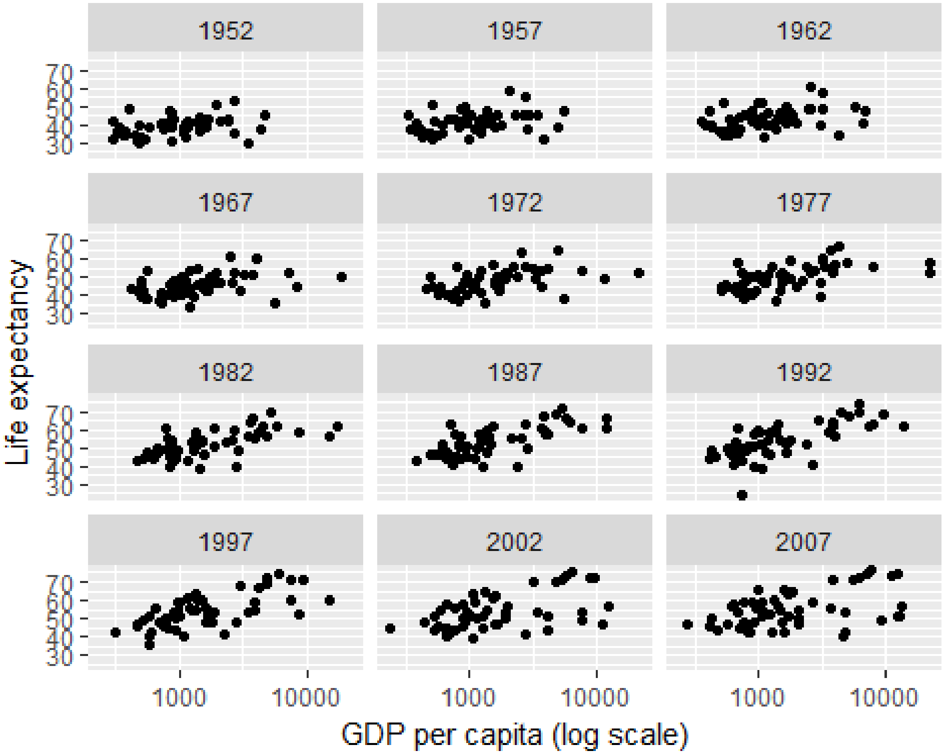


There are definetly differences in these distributions by continent. Europe has a tighter distribution and is center at a higher age. Africa has a more spread out distribution with few countries with as high of life expectancies as European countries. Asia is less spread out than Africa, but has a significant outlier. Oceana has sparce data.

Problem #2

(5 points.) For the gapminder data, make a subset consisting only of the data for African countries (for all years.) Using this subset, plot scatterplots with GDP per capita on the x-axis (on a log scale) and life expectancy on the y-axis, faceted by year. Describe in words how (i) GDP per capita in African countries, (ii) life expectancy in African countries, and (iii) the relationship between GDP per capita and life expectancy in African countries changed between 1952 and 2007.

```
gapminderAfrica = subset(gapminder, continent == "Africa")
gg = ggplot(gapminderAfrica, aes(x = gdpPercap, y = lifeExp)) + geom_point()
gg = gg + scale_x_log10() + xlab("GDP per capita (log scale)") + ylab("Life
expectancy")
gg = gg + facet_wrap(~ year, ncol = 3)
gg
```



(i) Looking at the data on a log scale it appears that GDP per capita has increased between the years of 1952 and 2007 for some of the African nations. This is particularly true from 1982 to 2007. There appears to be a cluster during that time period that saw increases in GDP per capita.

(ii) It looks like the average life expectency did increase for many African countries, particularly from 1952 to 1977. Post 1977 it looks like the increases were concentrated in the higher GDP nations.

(iii) It looks like life expectancy becomes more correlated with average GDP over time. In the years from 1952 to around 1972 there doesn't appear to be any slope on the plots. Starting in 1977 an upward slope begins to appear. The upward slope indicates a correlation between higher GDPs and longer life expectancies.

Problem #3

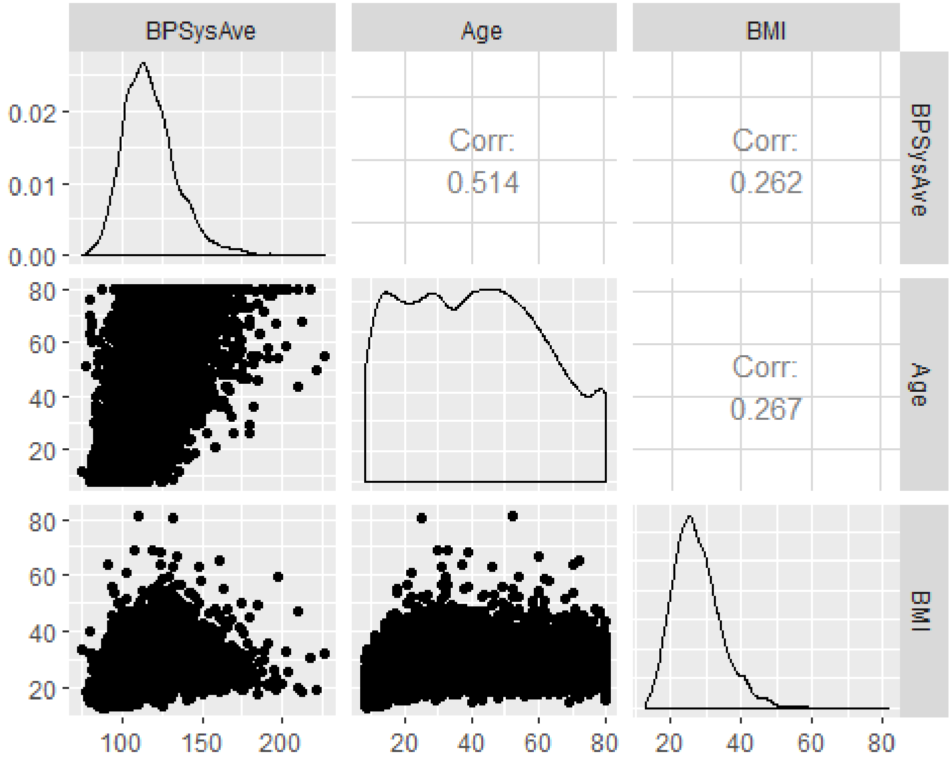
(15 points.) High systolic blood pressure is a strong predictor of heart attacks and strokes. A health researcher wants to know how average systolic blood pressure varies with age and body mass index (BMI.) For ease of interpretation, she does not wish to transform systolic blood pressure, but she is willing to consider interpretable transformations of the explanatory variables. She thinks the trends should be relatively smooth, but not necessarily linear. However, she is not interested in making predictions for individuals. She is not interested in formal inference right now, though she may be in the future. She knows some R, so you may include R code in your report, but she can't read your mind, so label your graphs.

Use the NHANES data set in the package of the same name to explore the researcher's questions. The relevant variables are: BPSysAve (the average of three measurements of systolic blood pressure, Age (in years; 80 or older is recorded as 80), BMI weight (in kilograms) divided by the square of height (in meters), Gender (male or female)

Draw graphs to show the relationship of average systolic blood pressure with (i) age and (ii) BMI. You should draw 2-4 graphs in total. For each graph you draw, include a brief justification of your modeling choices (type of model, transformations or lack of transformations) and verbally describe the trend, along with any differences you see between men and women. Some (sensibly rounded) quantitative measures may be useful, but you do not (and should not) list every single statistic you can think of.

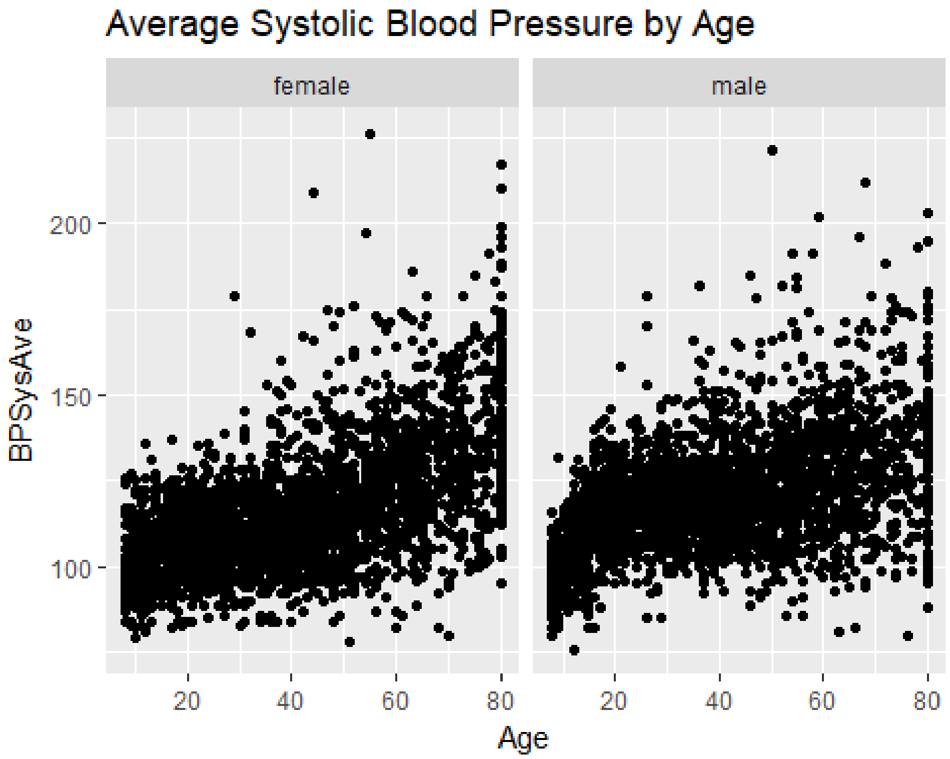
This graph show the relationship between the variables irregardless of age:

```
nhanes_sub = NHANES[c("BPSysAve", "Age", "BMI")]
nhanes_sub = na.omit(nhanes_sub)
ggpairs(nhanes_sub)
```



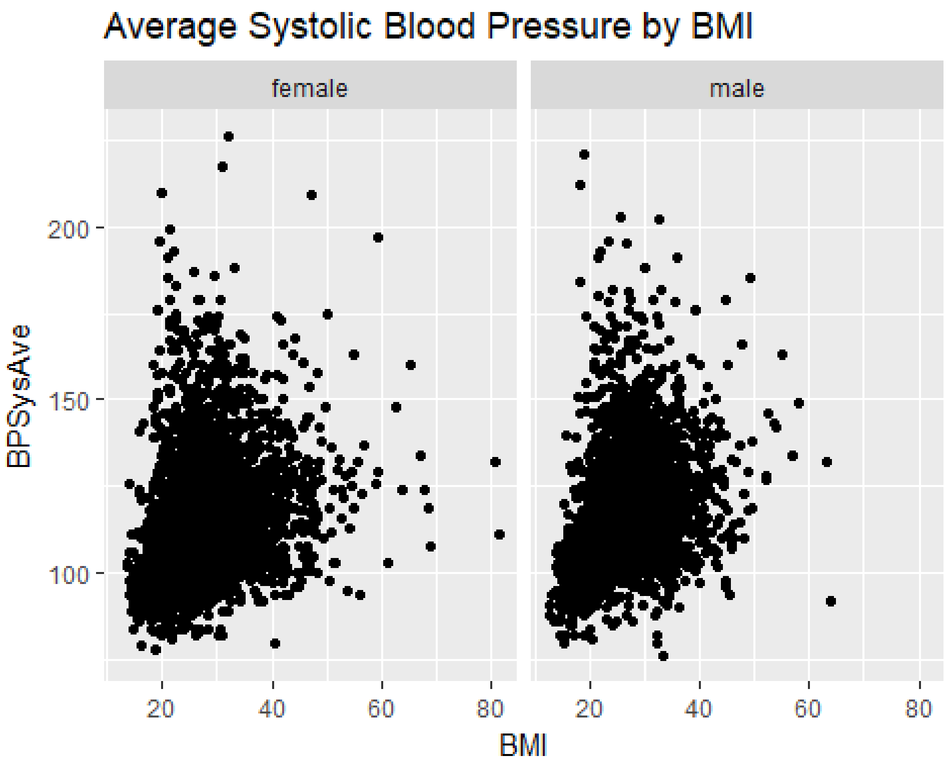
This graph shows the relationships between the variables irregardless of age. The strongest relationship appears to between the average sustolic blood pressure (SBP) and age. There appears to be a fairly linear relationship. The correlation coefficient between age and SBP is .514 which could be considered a moderate level of upward correlation. The relationship is between SBP and BMI is less obvious. It does appear there may be some linear relationship. The correlation coefficient between SBP and BMI is .262, which would be considered a low level of upward correlation.

```
nhanes_sub = NHANES[c("BPSysAve", "Age", "BMI", "Gender")]
nhanes_sub = na.omit(nhanes_sub)
gg = ggplot(nhanes_sub, aes(x=Age, y=BPSysAve))
gg = gg + geom_point() + facet_wrap(~Gender)
gg = gg + labs(title="Average Systolic Blood Pressure by Age")
gg
```



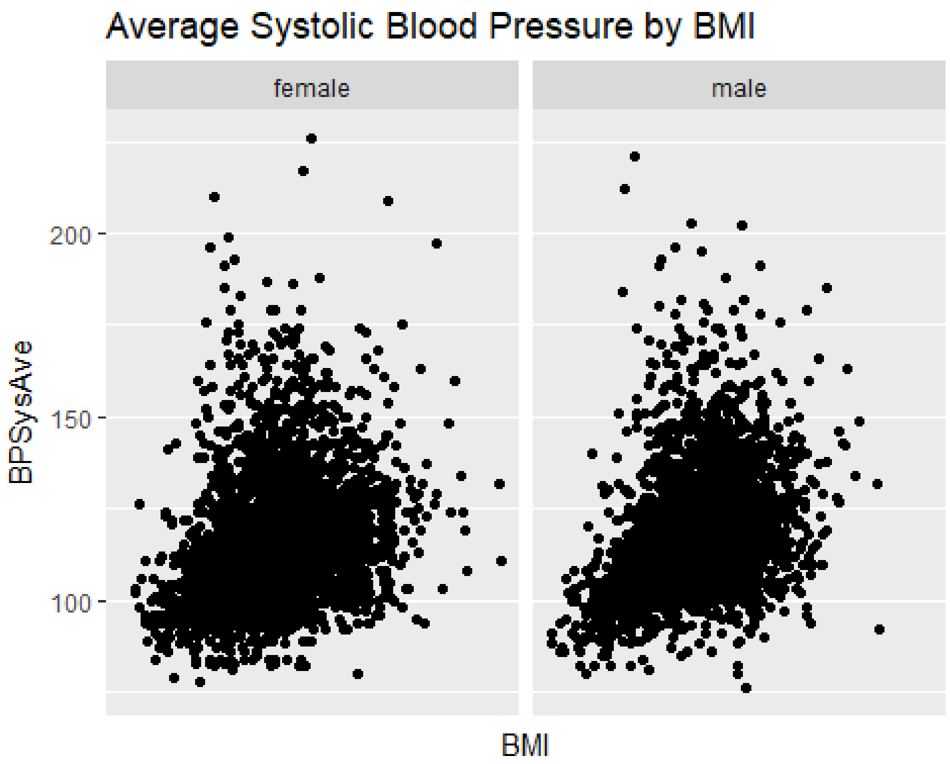
This graph shows the relationship between SBP and age, faceted by gender. There does appear to be a linear relationship across both genders. Though it does llok like for younger aged males the relationship seems to trend differently. The trend seems to icrease at a greater rate for the younger males.

```
nhanes_sub = NHANES[c("BPSysAve", "Age", "BMI", "Gender")]
nhanes_sub = na.omit(nhanes_sub)
gg = ggplot(nhanes_sub, aes(x=BMI, y=BPSysAve))
gg = gg + geom_point() + facet_wrap(~Gender)
gg = gg + labs(title="Average Systolic Blood Pressure by BMI")
gg
```



This graph shows the relationship between SBP and BMI, faceted by gender. The relationship here is not clear. It appears there may be some sort of ratlationship, but the data points seem to be failure centralized in a cluster. A relationship may become more clear if some sort of transformation were applied.

```r
nhanes_sub = NHANES[c("BPSysAve", "Age", "BMI", "Gender")]
nhanes_sub = na.omit(nhanes_sub)
gg = ggplot(nhanes_sub, aes(x=BMI, y=BPSysAve))
gg = gg + scale_x_log10() + geom_point() + facet_wrap(~Gender)
gg = gg + labs(title="Average Systolic Blood Pressure by BMI")
gg
```



This graph shows the relationship between SBP and BMI, faceted by gender with a log10 transformation to BMI. I chose this transformation becasue the cluster seemed to be grouped at lower BMIs. However with this transformation, there still doesn't appear to be a strong relationship between these variables.

Problem #4

(5 points.) We wish to study the heights of ten-year-olds using the NHANES data. The Height variable in that data frame gives heights in cm. Find:

(a) The mean height of the ten-year-olds in the sample;

```
dd = subset(NHANES, Age == 10)
dd = dd[c("Age", "Height")]
height_mean = mean(dd$Height)
height_mean

## [1] 144.1541
```

The mean is 144.1541 cm.

(b) The standard deviation of the heights of the ten-year-olds in the sample;

```
dd = subset(NHANES, Age == 10)
dd = dd[c("Age", "Height")]
height_sd = sd(dd$Height)
height_sd
```

```
## [1] 6.865524
```

The standard deviation is 6.85524 cm

(c) A 95% confidence interval for the mean height of all American ten-year-olds.

```
dd = subset(NHANES, Age == 10)
dd = dd[c("Age", "Height")]
height_mean = mean(dd$Height)
height_sd = sd(dd$Height)
height_cnt = length(dd$Height)
error <- qnorm(0.975)*height_sd/sqrt(height_cnt)
lower_bound = height_mean - error
upper_bound = height_mean + error
lower_bound

## [1] 142.8769

upper_bound

## [1] 145.4313
```

The 95% confidence interval is: lower bound = 142.8769 cm and the upper bound = 145.4313.

Problem #5

(5 points.) We measure the variables “TVs per 1000 people” and “Life expectancy” in a number of countries. We find the correlation between the two variables is 0.75, giving an r2 of about 0.56. However, most people would object to the statement that “The number of TVs per 1000 people explains about 56% of the variation in life expectancy between countries.” Explain why.

The problem with the statment: “The number of TVs per 1000 people explains about 56% of the variation in life expectancy between countries.” is that it asusmes that correlation is equivilaent to explanation or causation. In this case the fact that number of TVs correlates with life expectancy does not neccesarily mean that the numbers of TVs is driving or causing differences in life expectancy.

In this case the error is somewhat intuitive as the number of TVs driving life expectancy is counter-intuitive. One could conjecture that there are other drivers at play. For instance the the average number of TVs per household is likely very correlated to average income. It is much more intuitive to conjecture that average income could have a more casual relationship to life expectancy than simply the number of TVs.

So, fundamentally the statement “The number of TVs per 1000 people explains about 56% of the variation in life expectancy between countries.” confuses causation vs. correlation.

Problem #6

(5 points.) Consider the hypothesis test in https://xkcd.com/1132/: . The null hypothesis is that the sun has not gone nova; . The alternative hypothesis is that the sun has gone nova (i.e. exploded.)

(a) What is the Type I error rate of the frequentist statistician's test?

```
prob_2_sixes = (1/6)^2
prob_2_sixes

## [1] 0.02777778
```

Given the nature of the experiment, the type one error rate would be the probablity that the 'yes' answer is a lie. This is the probablity of rolling two sixes, or aproximately 2.78%.

(b) What is the power of the frequentist statistician's test?

```
prob_not_2_sixes = 1 - (1/6)^2
prob_not_2_sixes

## [1] 0.9722222
```

The test's power is the probability that the "yes" is true, which is the probablity of not rolling 2 sixes, or approximately 97.22%.

(c) The test appears to have a low Type I error rate (which is good) and high power (which is good.) So should the frequentist statistician take the Bayesian statistician's bet? If not, explain why not.

No, he should not take this bet. The frequentist test is not taking into account the probabilities associated with a nova actually occuring, where the Bayesian does. This can play out intuitively. If there is a one in a billion chance of a nova and the detector says that one has occurred with a ~97% accuracy, there is a far bigger chance that the double sixes had been rolled than that the nova actually occurred.

Problem #7

(5 points.) Suppose we test the null hypothesis that a coin is fair by tossing it 50 times. We decide that we will reject the null hypothesis if we see either 33 or more heads, or 33 or more tails.

Ho: p = .5 H1: p != .5

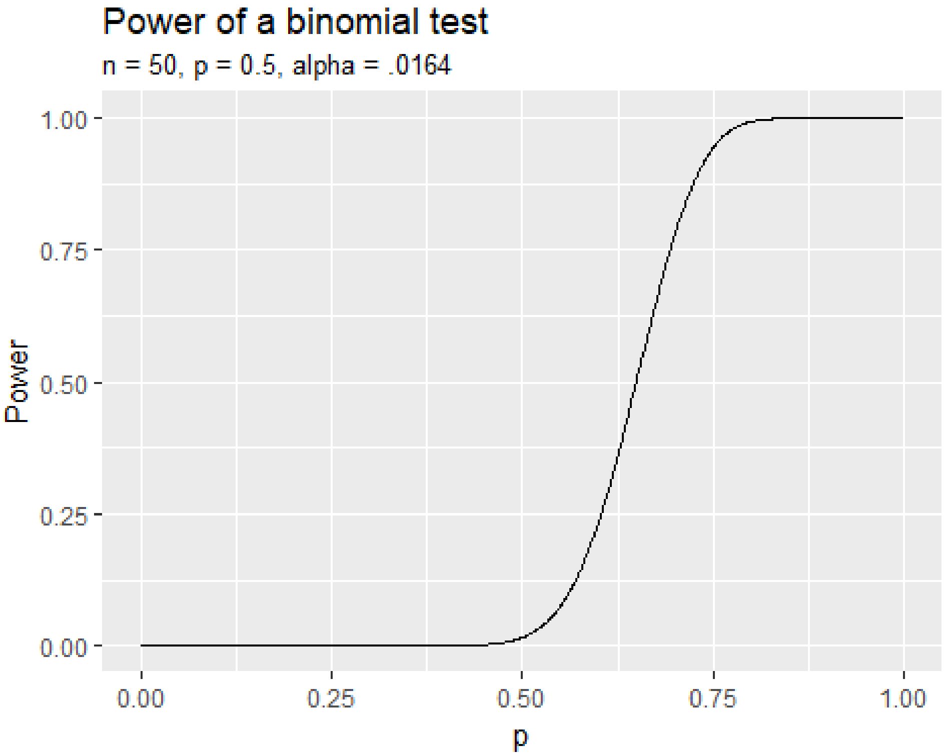
(a) What is the true significance level of this test?

```
prob_33_or_greater = 1 - pbinom(32, 50, .5)
prob_33_or_greater

## [1] 0.01641957
```

The true significance is the probabilty of getting 33 or more heads given that the coin is fair. This is 1 minus the probablity of getting 32 or less heads given the coin is fair, or approximatley 1.64%

(b) Plot the power of this test as a function of p, the probability of getting heads.

```
p = seq(0, 1, .001)
Power = 1 - pbinom(32, 50, p)
power.df = data.frame(p, Power)
gg = ggplot(power.df, aes(x = p, y = Power)) + geom_line()
gg = gg + labs(title = "Power of a binomial test", subtitle = "n = 50, p =
0.5, alpha = .0164")
gg
```



(c) How extreme does p have to be to get 80% power from this test?

```
p = seq(0, 1, .001)
Power = 1 - pbinom(32, 50, p)
power.df = data.frame(p, Power)
power.df[min(which(power.df[,2]>=.8)),]

##         p     Power
## 705 0.704 0.8003199
```

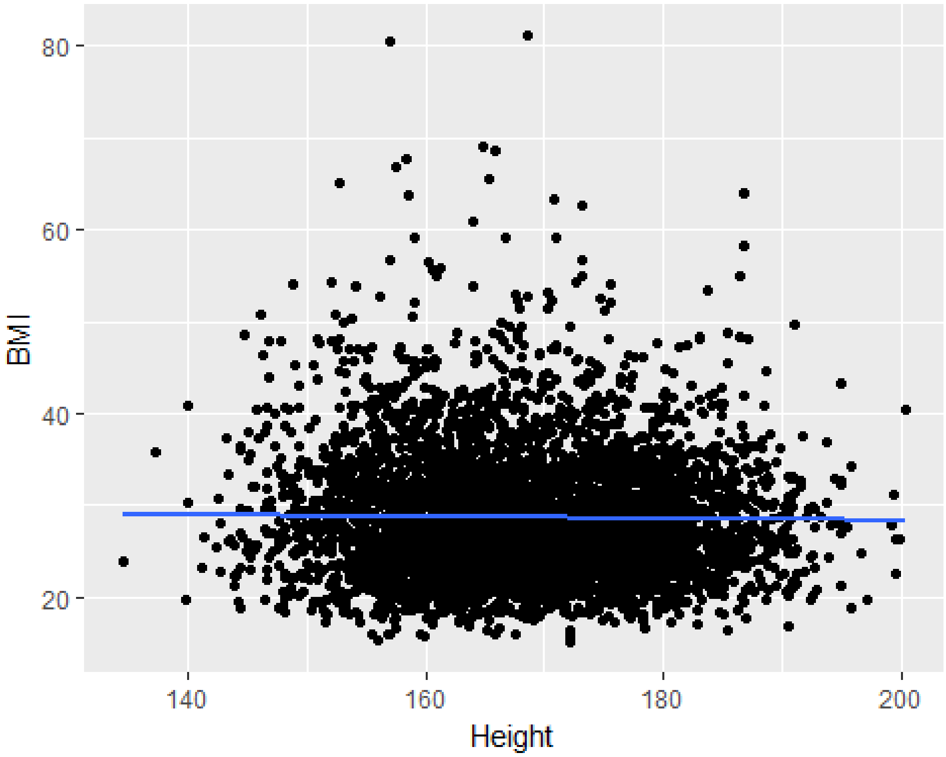
In order to get an 80% power p must be approximatley .704.

Problem #8

(5 points.) For the NHANES data, we wish to examine whether expected BMI varies as a function of Height for adults (people aged at least 18.)

(a) Using the lm() function or otherwise, fit a simple linear regression with Height as the x-variable and BMI as the y-variable. Find a P-value for a test of the null that the coefficient of Height in this model is zero, and explain what this P-value tells you about height and BMI.

```
dd = subset(NHANES, Age >= 18)
dd = dd[c("BMI", "Height")]
dd = na.omit(dd)
ggplot(dd, aes(x = Height, y = BMI)) + geom_point() + geom_smooth(method =
"lm", se = FALSE)
```



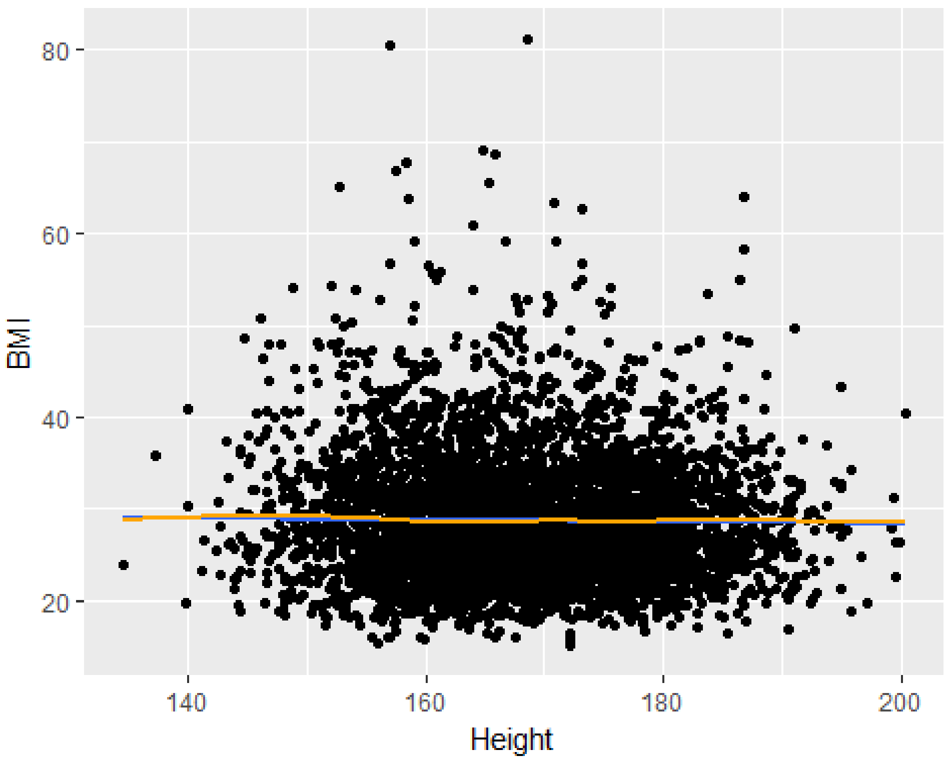
```
summary(lm(BMI ~ Height, data = dd))

##
## Call:
## lm(formula = BMI ~ Height, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.622  -4.730  -1.044   3.406  52.574
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 30.238002   1.302816  23.210   <2e-16 ***
## Height      -0.009271   0.007702  -1.204    0.229    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.698 on 7412 degrees of freedom
## Multiple R-squared:  0.0001955,  Adjusted R-squared:  6.058e-05 
## F-statistic: 1.449 on 1 and 7412 DF,  p-value: 0.2287
```

The simple linear regression of Height to BMI produces an intercept of approximately 30.238 and a slop of approximatley -.009. The p-value for the null hypothesis that the coefficient of Height in this model is zero is 0.2287.The p-value is a measure of the probablity of the test statistic being this unusual or more, given the null hypothesis was correct. With a p-value of .2287, this is a fairly large p-value and generally would not lead to rejecting the null hypothesis. Practicaly this means there is some reason to think that height is not a good predictor of BMI.

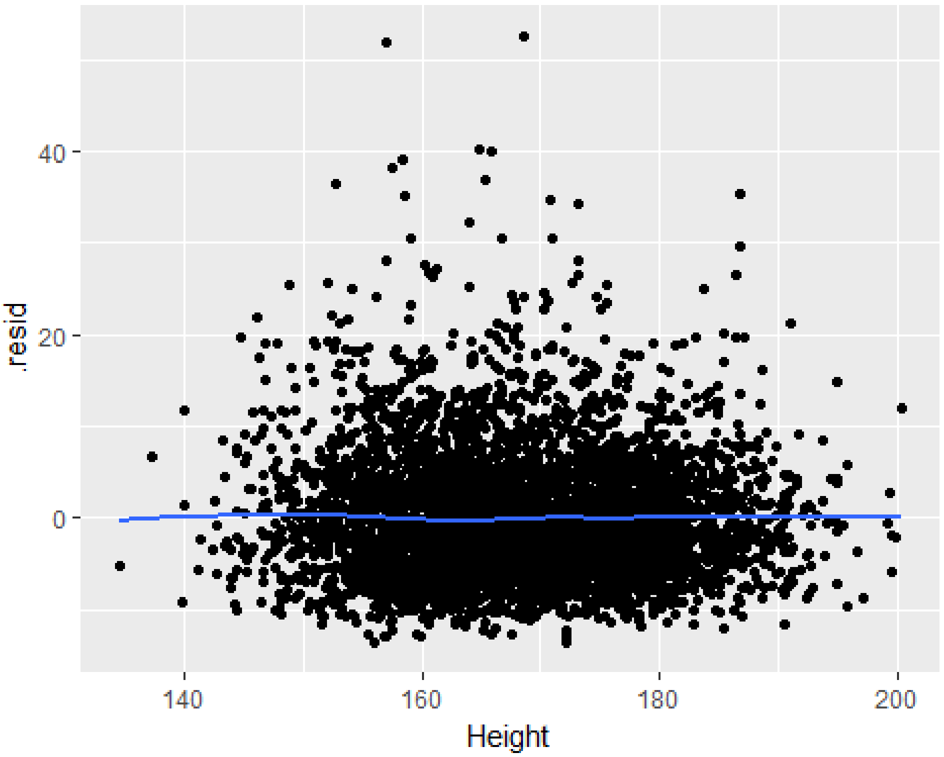
(b) Using graphs and context, check the assumptions of your regression (linearity, independence, equal variance of errors, normality of errors.) Are the assumptions required for your P-value in (a) to be valid approximately met?

```r
dd = subset(NHANES, Age >= 18)
dd = dd[c("BMI", "Height")]
dd = na.omit(dd)
ggplot(dd, aes(x = Height, y = BMI)) + geom_point() + geom_smooth(method =
"lm", se = FALSE) + geom_smooth(method = "loess", se = FALSE, color =
"orange")
```



Checking for linearity, I compare the linear fit ot a loess fit. In this case they are very close, which is an indicator of linearity.

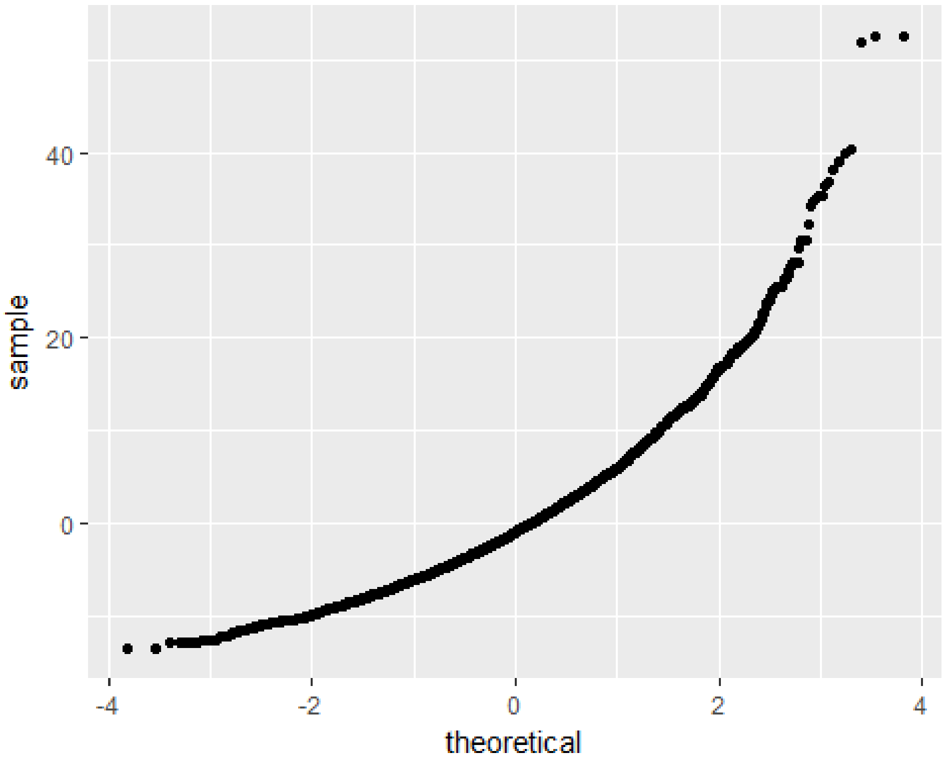
```
dd = subset(NHANES, Age >= 18)
dd = dd[c("BMI", "Height")]
dd = na.omit(dd)
dd.lm = lm(BMI ~ Height, data = dd)
dd.lm.df = augment(dd.lm)
ggplot(dd.lm.df, aes(x = Height, y = .resid)) + geom_point() +
geom_smooth(method = "loess", se = FALSE)
```



This is a graph of the residuals of the linear regression. There is some slight curvature to the loess curve, but overall there does not seem to be a trend. This would indicate the model is reasonably correct in capturing the trend.

There does seem to be some changes in variance as you move from left ot right of the graph. This may be a slight violation of heteroscedasticity. The points also do not seem to be symetric. there seems to be a higher spread above the zero line than below.

```
dd = subset(NHANES, Age >= 18)
dd = dd[c("BMI", "Height")]
dd = na.omit(dd)
dd.lm = lm(BMI ~ Height, data = dd)
dd.lm.df = augment(dd.lm)
ggplot(dd.lm.df, aes(sample = .resid)) + stat_qq()
```



This qq plot is a check on the normality of the residuals. In this case there is a definite curve which is an indicator that the residuals are not normally destributed.