Stats 10 Lab 3
Name: Brandon Truong

Section 1

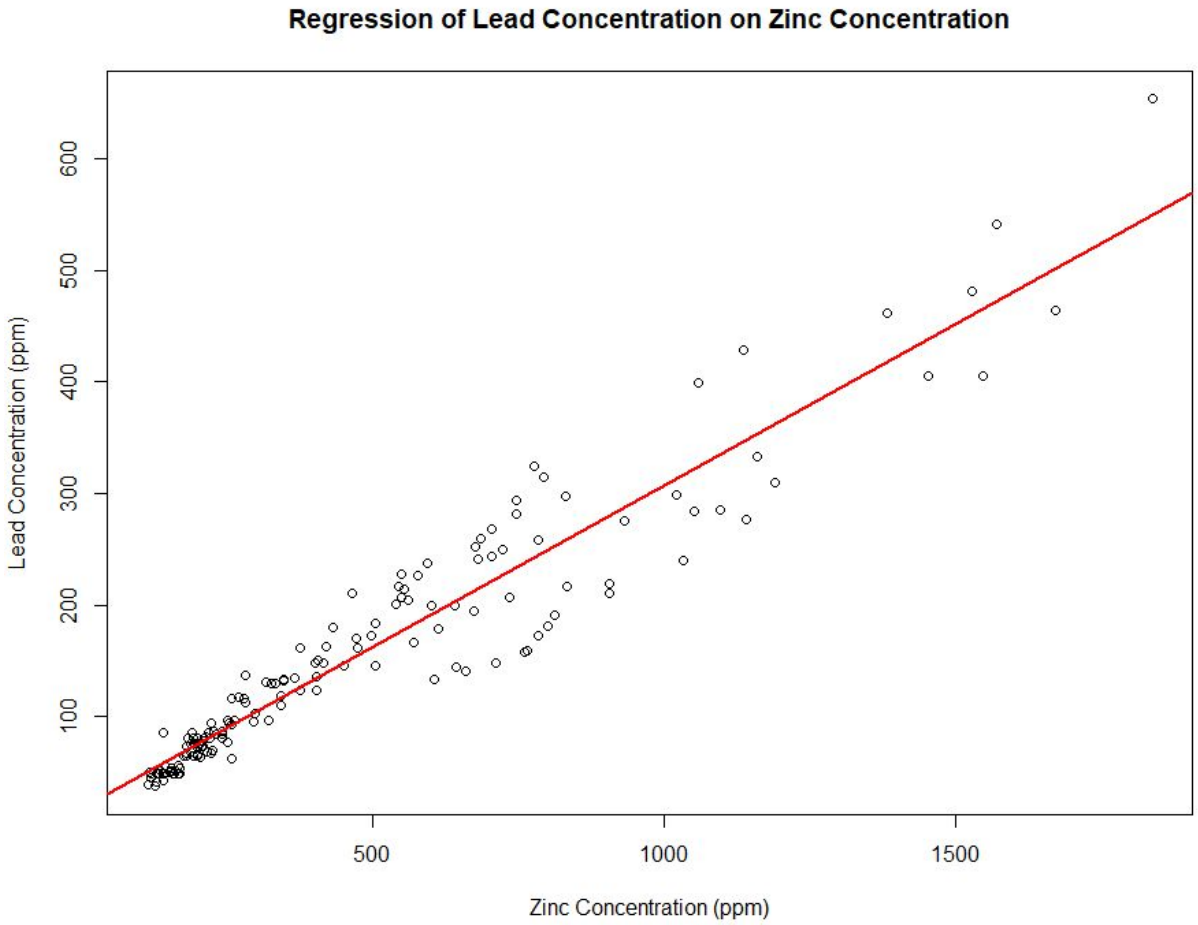
```
soil <-
read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/soil_complete.txt", header=TRUE)
```

a)

```
> linear_model <- lm(soil$lead ~ soil$zinc)
> summary(linear_model)

Call:
lm(formula = soil$lead ~ soil$zinc)

Residuals:
    Min      1Q  Median      3Q     Max
-79.853 -12.945  -1.646  15.339 104.200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.367688   4.344268   3.998 9.92e-05 ***
soil$zinc    0.289523   0.007296  39.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.24 on 153 degrees of freedom
Multiple R-squared:  0.9114,    Adjusted R-squared:  0.9109
F-statistic:  1575 on 1 and 153 DF,  p-value: < 2.2e-16
```

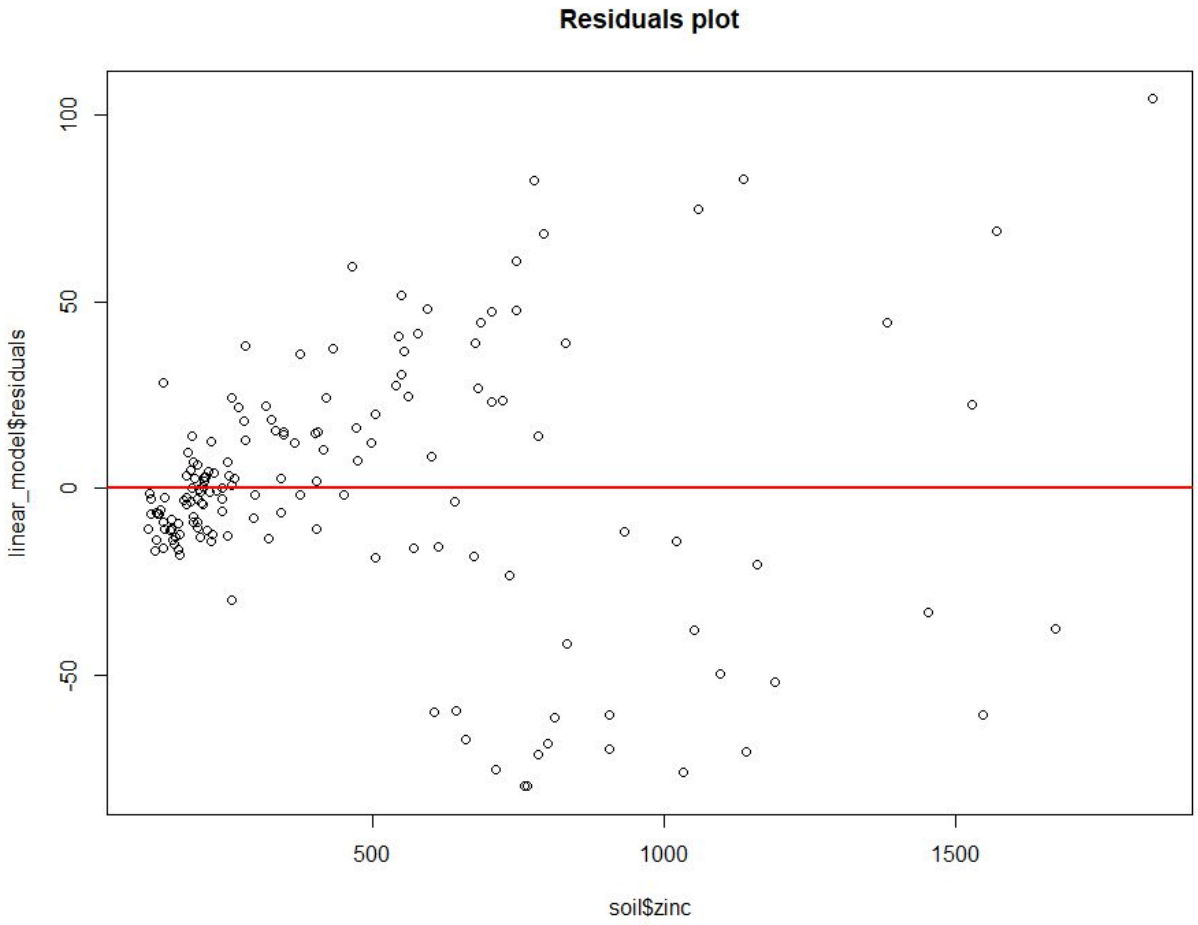
b)

```
> plot(soil$lead ~ soil$zinc, xlab = "Zinc Concentration (ppm)", ylab = "Lead Concentration (ppm)", main = "Regression of Lead Concentration on Zinc Concentration", )
> abline(linear_model, col = "red", lwd = 2)
```



c)

```
> plot(linear_model$residuals ~ soil$zinc, main = "Residuals plot")
> abline(a = 0, b = 0, col = "red", lwd = 2)
```



d)
The equation of the regression line is:

[predicted] lead ppm = 17.37 + 0.29 * zinc ppm

e)
The predicted lead ppm is 307.37 ppm
> 17.37 + 0.29 * 1000 = predicted lead ppm
[1] 307.37

f)
We predict the lead concentration difference (ppm) to be 29 ppm
> 0.29 * 100 = difference in lead ppm (from A to B)
[1] 29

g)
From summary of linear_model:
91.14% of the variability in lead is explained by the variability in

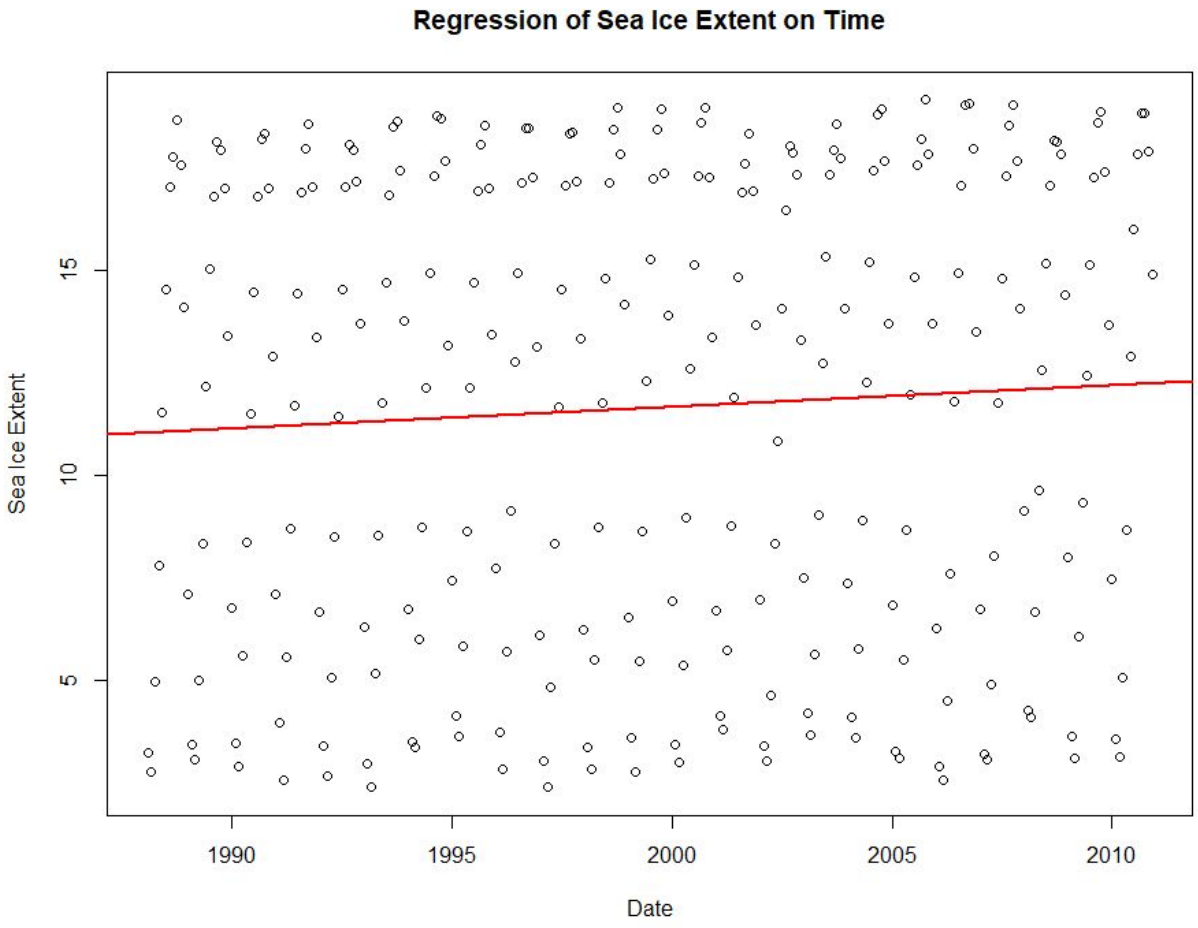
zinc
> 0.9114
[1] 0.9114

h)
Of the three main assumptions of linear regression, we believe that linearity and symmetry are satisfied. We can see this in the regression scatterplot, which shows a linear fit that suggests symmetry between the two variables. This means that exchanging independent and dependent variables should not affect r^2. However, the variance is not constant, since the residual plot shows a slight fan shape, suggesting possible heteroscedasticity.
> Linearity okay
> Symmetry okay
> Equal Variance: Variance is not constant

Section 2
```
> ice <- read.csv('~/UCLA Coursework/STATS 10/sea_ice.csv', header = T)
> colnames(ice)
[1] "Date"   "Extent"
> ice$Date <-as.Date(ice$Date, "%m/%d/%Y")
```
a)
```
> linear_model <-lm(ice$Extent~ice$Date)
> summary(linear_model)

Call:
lm(formula = ice$Extent ~ ice$Date)

Residuals:
   Min     1Q Median     3Q    Max
-9.445 -5.439  1.442  5.599  7.564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.011e+01  1.558e+00   6.486 4.11e-10 ***
ice$Date    1.438e-04  1.411e-04   1.019    0.309
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.654 on 273 degrees of freedom
Multiple R-squared:  0.003787,  Adjusted R-squared:  0.0001377
F-statistic: 1.038 on 1 and 273 DF,  p-value: 0.3093
```

b)
There appears to be no trend in the data since the r^2 variance is low (0.003787), which we can tell by the almost horizontal data, suggesting that sea ice extent is not affected by time

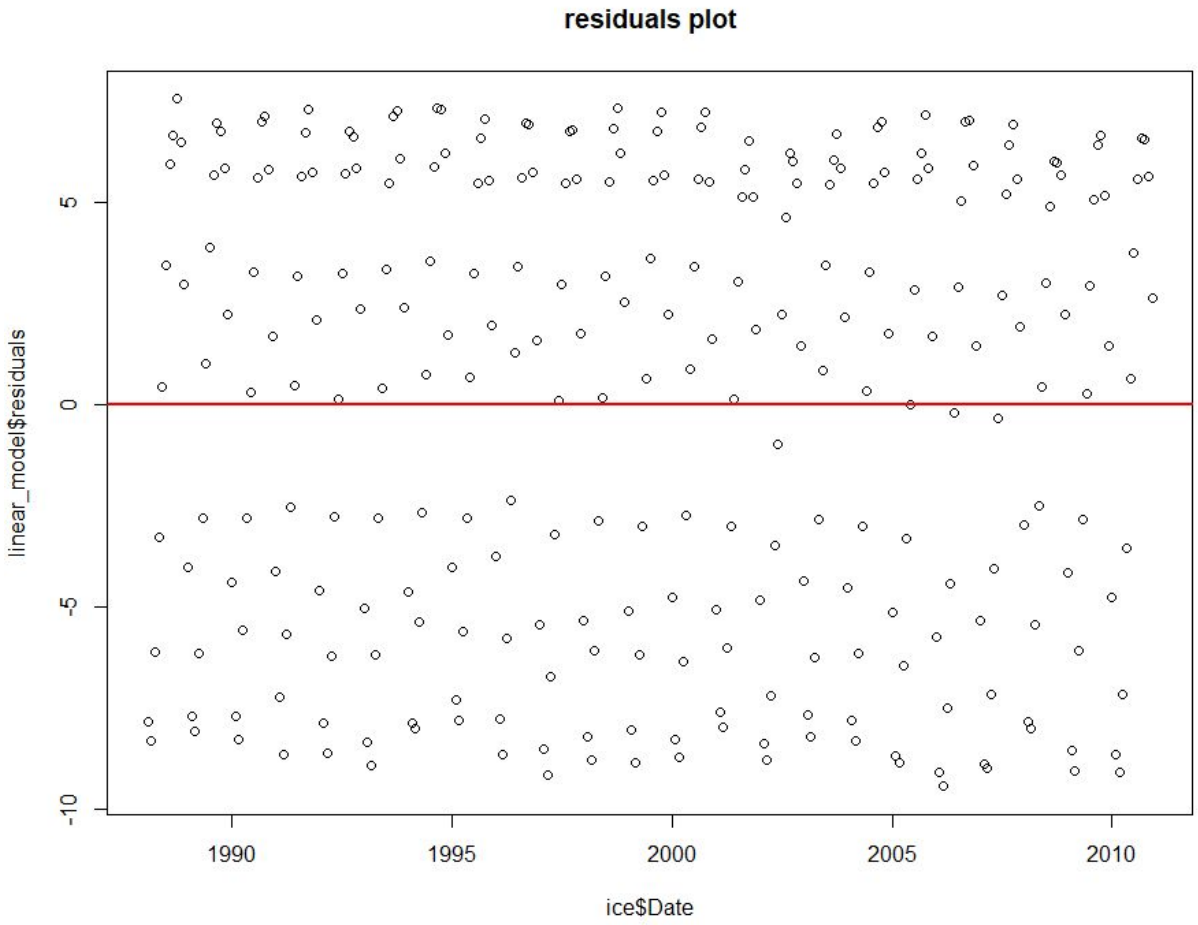
```
> plot(ice$Extent~ice$Date, ylab = "Sea Ice Extent", xlab = "Date",
main = "Regression of Sea Ice Extent on Time")
> abline(linear_model, col = "red", lwd = 2)
```




```
> #no trend
```

c)
We should be worried about linearity, since the data suggests no linear fit due to the variables not appearing to be related to one another

```
> plot(linear_model$residuals ~ ice$Date, main = "residuals plot")
> abline(a = 0, b = 0, col ="red", lwd = 2)
```



Section 3

a)

These are the possible rolls that win

doubles money : (1,6),(6,1),(2.5),(5,2),(3,4),(4,3),(5,6),(6,5)

8/(6*6)=2/9

These are the possible rolls that lose

loses all : (1,1),(1,2),(2,1),(6,6)

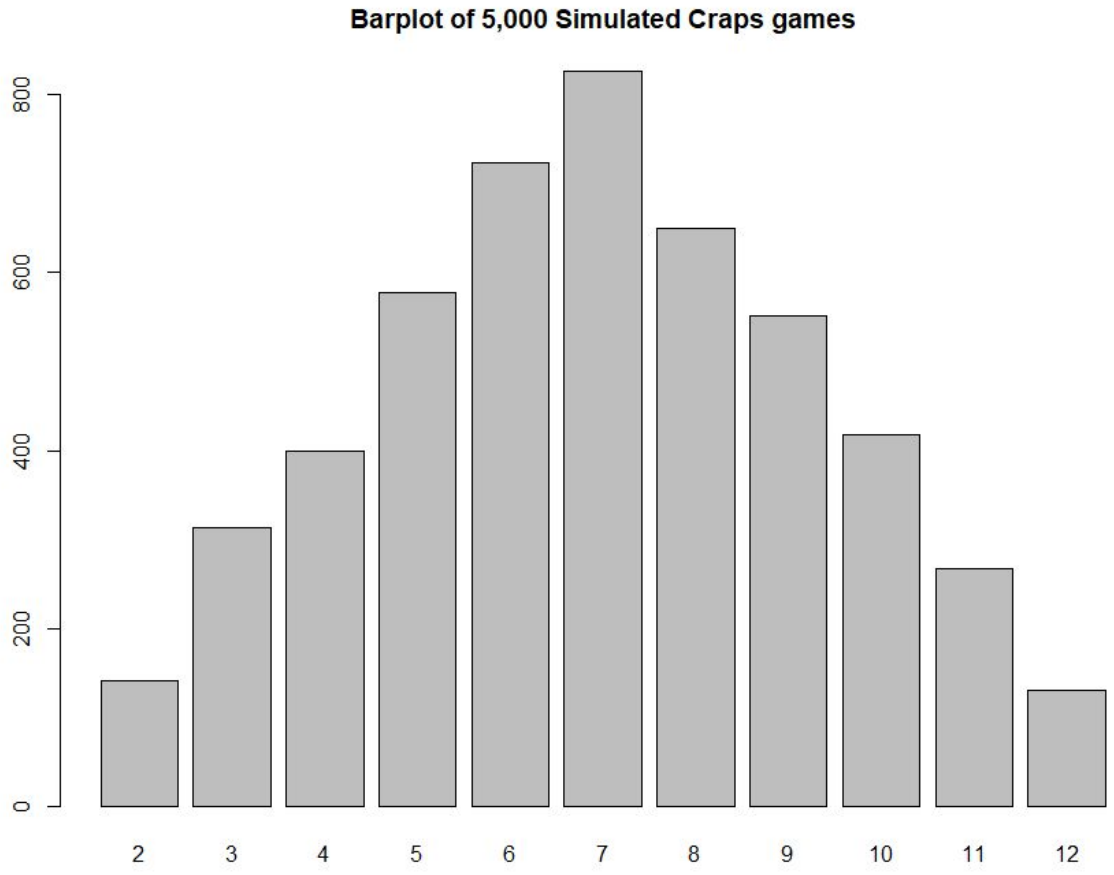
4/(6*6)=1/9

Since there are a total of 6*6=36 possible rolls, that means that the possibility of winning is 8/36=2/9 and the possibility of losing is 4/36=1/9

b)

```
set.seed(123)
> numbers = 1:6
> rand_draws = replicate(5000, sample(numbers, 2, replace = TRUE))
```

```
> results = colSums(rand_draws)
> table(results)
results
  2   3   4   5   6   7   8   9  10  11  12
141 314 400 577 723 826 650 552 418 268 131

> barplot(table(results), main = "Barplot of 5,000 Simulated Craps
games")
```



c)

Adam doubles his money 21.88% of the time and loses all his money 11.72% of the time.

```
> #P(doubles money)
> sum(table(results)[c(7,11)-1]/5000)
[1] 0.2188
> #P(loses all)
> sum(table(results)[c(2,3,12)-1]/5000)
[1] 0.1172
```

d)

The events of Adam winning and losing are disjoint and not independent. This is because Adam can not win and lose at the same time. The events are not independent since if we know that Adam won his game, it also tells us that he did not lose either. This means that if we know that if Adam winning happened, then our understanding of the probability of Adam losing is changed. This is supported by the fact that disjoint events can not be independent

e)
We can determine that the events are not independent if we use the independent definition Prob(A|B)= Prob(A) which tells us that knowledge of one event does not affect the probability of the other

Prob(Winning (A)) = 2/9
Prob(Losing (B)) = 1/9
prob(Winning and Losing (A & B)) = 0, according to disjoint

Since Prob(A|B) = Prob(A & B) / Prob(B) and Prob(A|B) = Prob(A) if it's independent, then therefore

Prob(A) = Prob(A & B) / Prob(B)

However, we get
2/9 =/= 0 / (1/9)

Since Prob(A|B)=/=Prob(A), we can conclude they are not independent.

Section 4
a)
```
> n=365
> p=0.40
```

b)
The mean is 146 days and the standard deviation is 9.359487 days
```
> n*p #mean
[1] 146
> sqrt(n*p*(1-p)) #standard deviation
[1] 9.359487
```

c)
The probability of heavy rain for exactly 145 days is 0.04239996 or 4.239996%
```
> dbinom(145,size=n,prob=p)
```

```
[1] 0.04239996
```

d)
The probability of between 125 and 175 days of heavy rain is 0.9888137 or 98.88137%

```
> pbinom(175,size=n,prob=p)-pbinom(124,size=n,prob=p)
[1] 0.9888137
```

e)
The probability of more than 230 inches of rain is 0.0668072 or 6.68072%

```
> 1-pnorm(230,mean=200,sd=20)
[1] 0.0668072
> pnorm(230,mean=200,sd=20,lower.tail=F)
[1] 0.0668072
```