Trent Bellinger

Stats 10 Lab 3

Exercise 1

soil <- read.table('soil_complete.txt', header = TRUE)

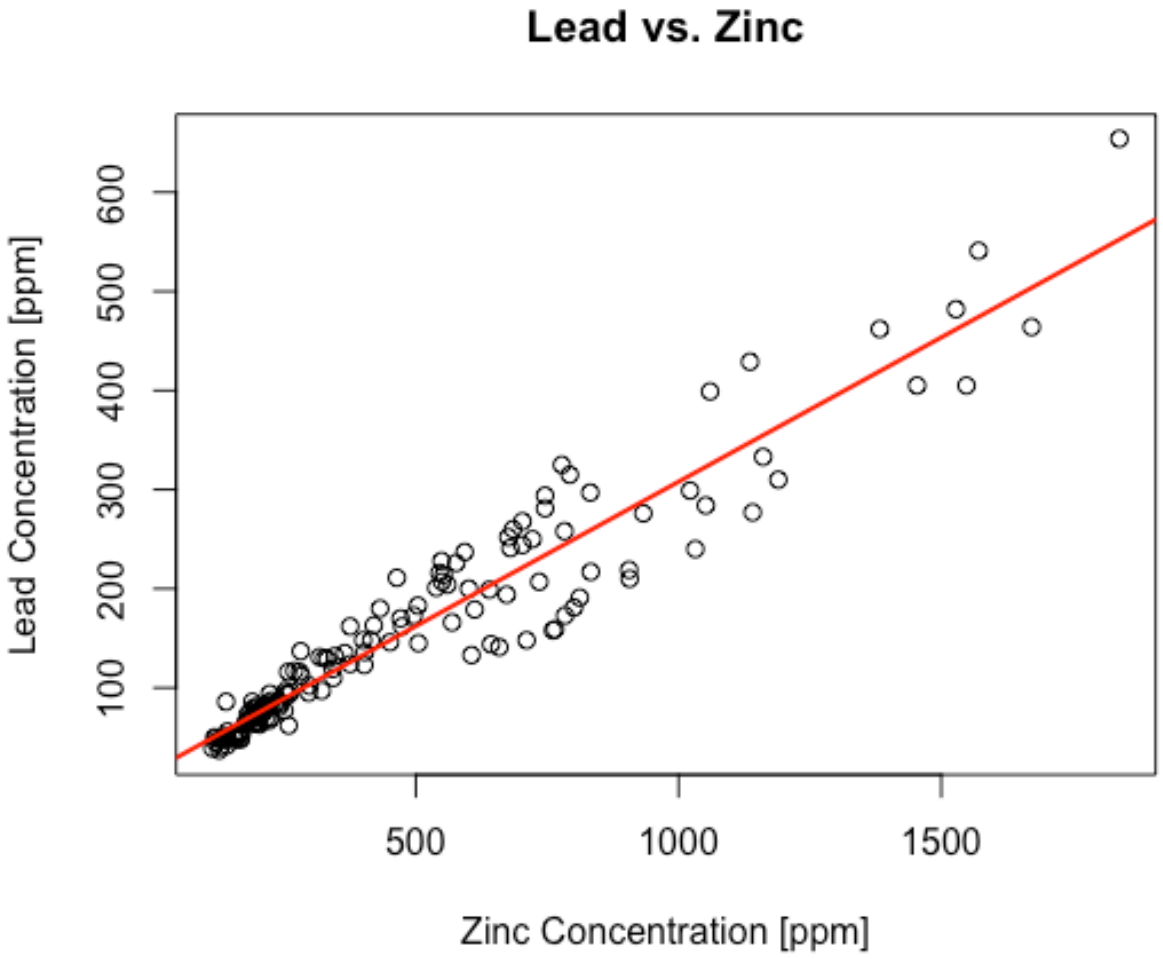
1a

soil_model <- lm(lead~zinc, data=soil)
summary(soil_model)

```
Call:
lm(formula = lead ~ zinc, data = soil)

Residuals:
   Min      1Q  Median      3Q     Max
-80.455 -12.570  -1.834  15.946 101.651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.582928   4.410443    3.76 0.000244 ***
zinc         0.291335   0.007415   39.29  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.37 on 149 degrees of freedom
Multiple R-squared:  0.912,    Adjusted R-squared:  0.9114
F-statistic:  1544 on 1 and 149 DF,  p-value: < 2.2e-16
```

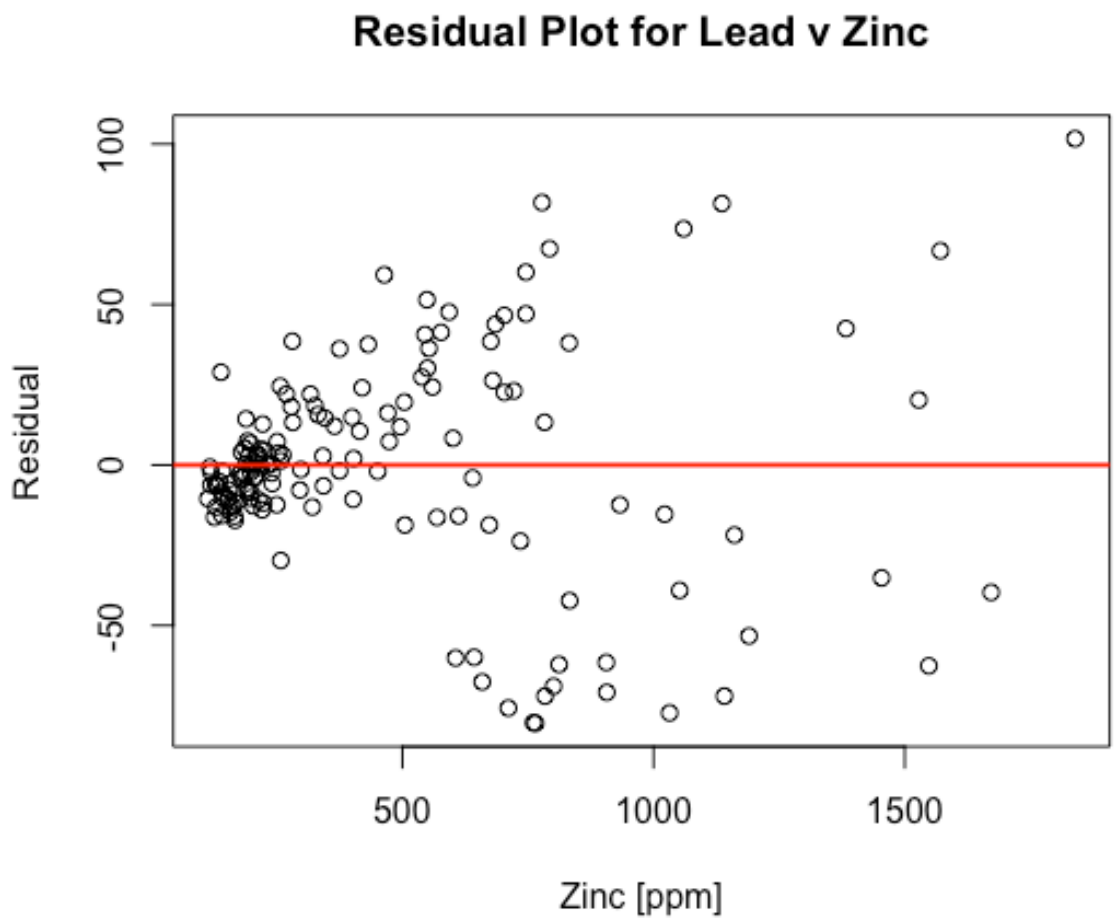
1b

```
plot(lead~zinc, data=soil, xlab = 'Zinc Concentration [ppm]', ylab = 'Lead Concentration
        [ppm]', main = 'Lead vs. Zinc')
abline(soil_model, col = 'red', lwd = 2)
```



1c

```
plot(soil_model$residuals~soil$zinc, xlab = 'Zinc [ppm]', ylab = 'Residual',
     main = 'Residual Plot for Lead v Zinc')
abline(0, 0, col = 'red', lwd = 2)
```



1d

y = ax + b
lead = 0.29*zinc + 16.58

1e

predict(soil_model, data.frame(zinc = 1000))

307.9184

The lead concentration at this point is expected to be 307.9184.

1f

soil_coeffs <- soil_model$coefficients
soil_coeffs[2] * 100

zinc
29.13355

We expect the lead concentration in location A to be 29.13355 ppm higher in location A compared to location B.

1g

summary(soil_model)

Multiple R-Squared: 0.912, Adjusted R-Squared: 0.9114

cor(soil$zinc, soil$lead)^2

[1] 0.9119854

About 91.2% of the variation in lead can be explained by the variation in zinc.

1h

Linearity of the data is met as shown in subpart (b).
Symmetry is not respected as the residuals vary from -50 to 100, which is not symmetric.
Equal variance assumption is violated as we see a clear fan-shape of x-dependence between the residual and the x-variable.
This is not ignorable since the points above the x-axis and below the x-axis are of similar density. The regression line is less accurate as the concentrations increase, so it is not a perfect fit.

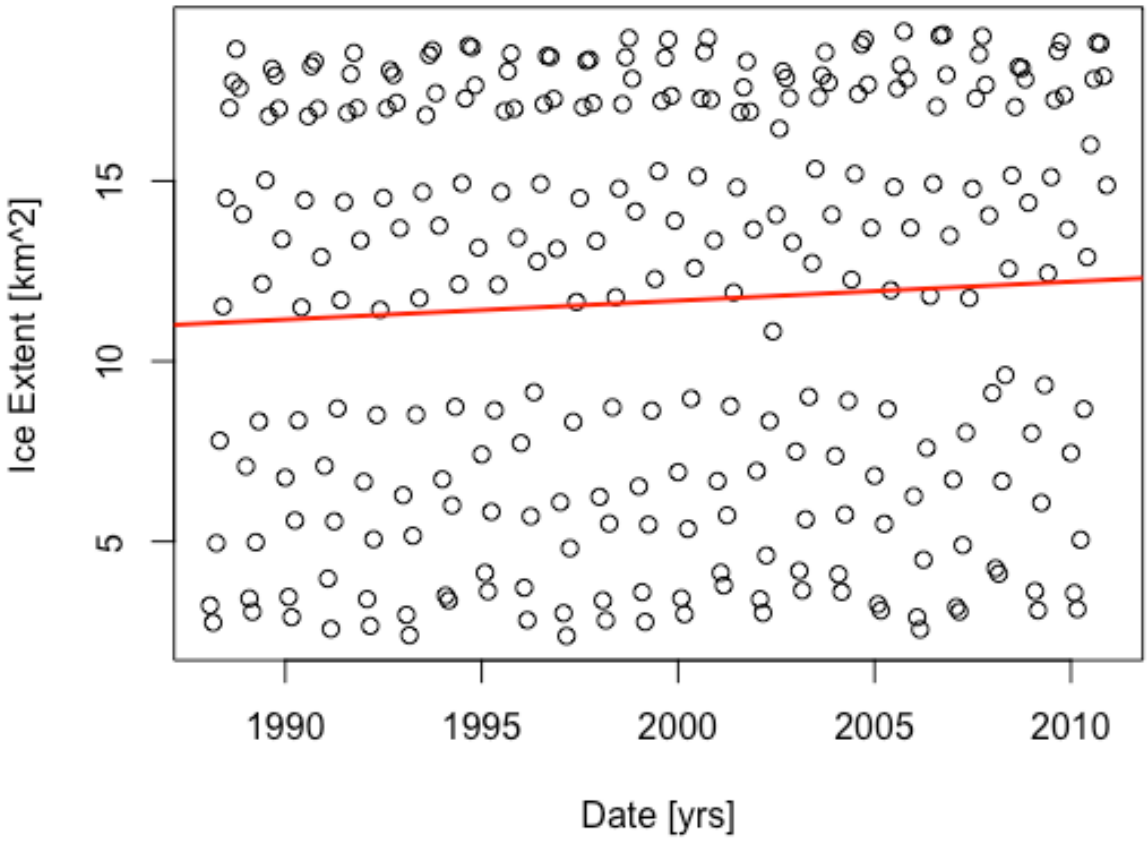
Exercise 2

```
ice <- read.csv("sea_ice.csv", header = TRUE)
ice$Date <- as.Date(ice$Date, "%m/%d/%Y")

# 2a

ice_model <- lm(Extent~Date, data = ice)
summary(ice_model)
```

```
Call:
lm(formula = Extent ~ Date, data = ice)

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 ***
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
```

2b

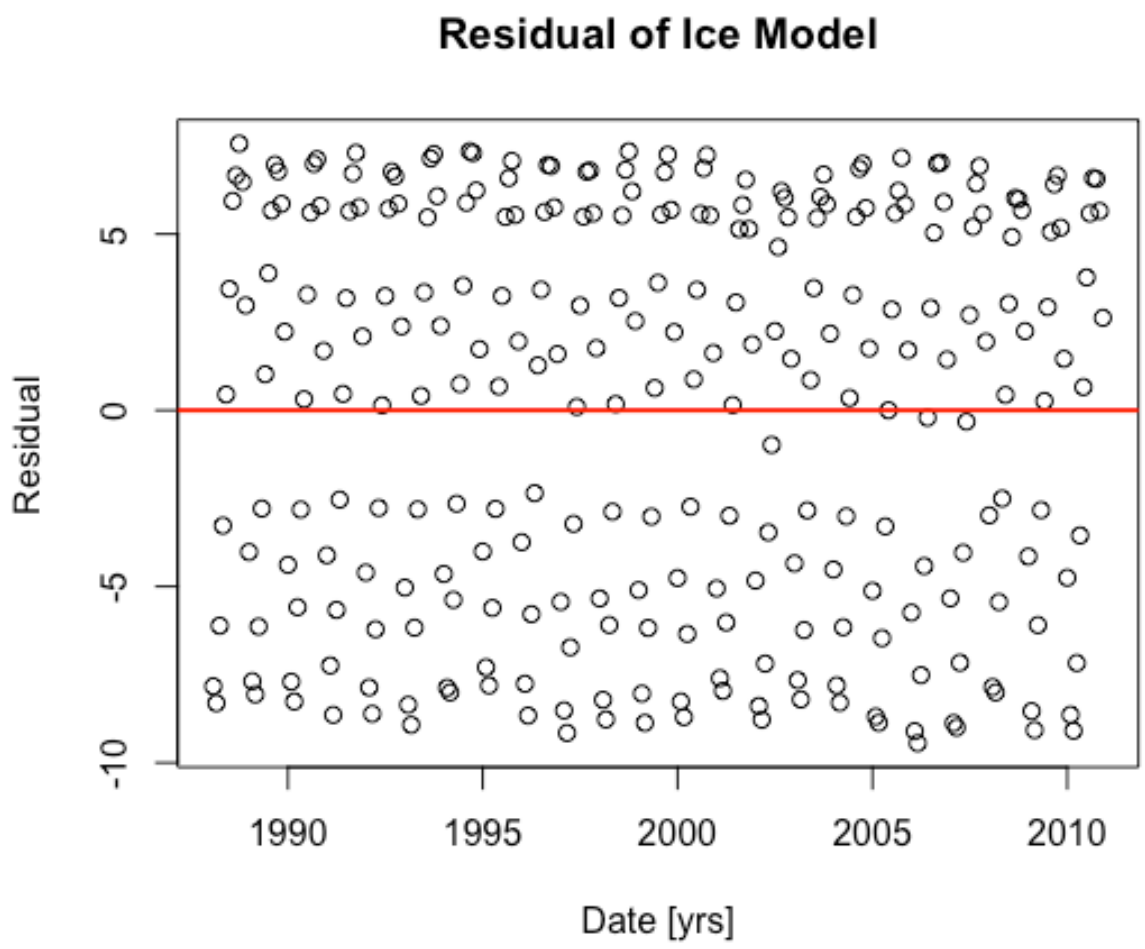
plot(Extent~Date, data = ice, xlab = "Date [yrs]", ylab = "Ice Extent [km^2]")
abline(ice_model, col = 'red', lwd = 2)



The data does show a slight positive trend according to the regression line.

2c

```
plot(ice_model$residuals~ice$Date, xlab = "Date [yrs]", ylab = "Residual",
    main = "Residual of Ice Model")
abline(0, 0, col = 'red', lwd = 2)
```



There are two main issues. The first issue is that the data is not linear as can be seen in part 2(b). The second Issue is that there seems to be a lack of x-symmetry as the residuals range from -10 to 5. Thus, the regression line is not an accurate representation of the data.
However, the equal variance assumption is respected. There is no x-dependence on residual variability.

Exercise 3

3a

Total options: 6*6 = 36

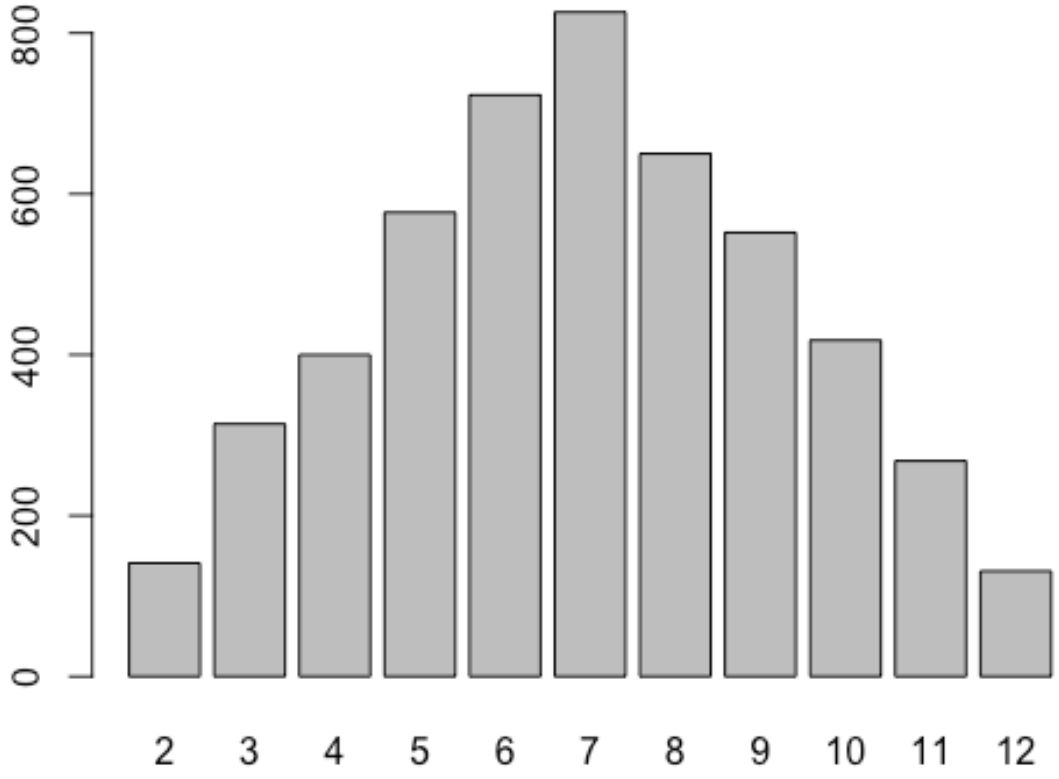
Ways to sum to 7: 3+4, 4+3, 5+2, 2+5, 6+1, 1+6 => 6 options
Ways to sum to 11: 6+5, 5+6 => 2 options
P(win) = 8/36 = 2/9

Ways to sum to 2: 1+1 => 1 option
Ways to sum to 3: 1+2, 2+1 => 2 options
Ways to sum to 12: 6+6 => 1 option
P(loss) = 4/36 = 1/9

The chance that Adam will double his money in the first round is 2/9.
The chance that Adam will lose his money in the first round is 1/9.

3b

```
set.seed(123)
dice <- 1:6
outcomes <- replicate(5000, sample(dice, 2, replace=TRUE))
roll_sums <- colSums(outcomes)
barplot(table(roll_sums))
```



#3c

prW <- mean((roll_sums == 7)|(roll_sums == 11))
prW

[1] 0.2188

prL <- mean((roll_sums == 2)|(roll_sums == 3)|(roll_sums == 12))
prL

[1] 0.1172

The percentage of time Adam doubles his money was 21.88%.
The percentage of time Adam lost his money was 11.72%.

3d

Independence: P(A and B) = P(A) * P(B) or P(A|B) = P(A) => P(A|B) = P(A and B)/P(B) = P(A)
Disjoint: P(A and B) = 0

Adam winning money and losing money are disjoint events because both can not happen at the same time. Therefore, they cannot be independent as well since either winning or losing has a non-zero probability of happening.

3e

W_event <- (roll_sums == 7)|(roll_sums == 11)
L_event <- (roll_sums == 2)|(roll_sums == 3)|(roll_sums == 12)
prWL <- mean(W_event & L_event)

prWL

[1] 0

prWL != (prW * prL)

[1] TRUE

The probability of a win and a loss happening is not equal to the probability of a win times the probability of a loss, so the events are not independent.